Improving Clinical Risk-Stratification Tools: Instance-Transfer for Selecting Relevant Training Data

by

Jen J. Gong

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Author .... Signature redacted ....
Department of Electrical Engineering and Computer Science
May 22, 2014

Certified by.. Signature redacted ....
John V. Guttag
Dugald C. Jackson Professor
Thesis Supervisor

Accepted by .... Signature redacted ....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Improving Clinical Risk-Stratification Tools: Instance-Transfer for Selecting Relevant Training Data

by

Jen J. Gong

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2014, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

One of the primary problems in constructing risk-stratification models for medical applications is that the data are often noisy, incomplete, and suffer from high class-imbalance. This problem becomes more severe when the total amount of data relevant to the task of interest is small. We address this problem in the context of risk-stratifying patients receiving isolated surgical aortic valve replacements (isolated AVR) for the adverse outcomes of operative mortality and stroke. We work with data from two hospitals (Hospital 1 and Hospital 2) in the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database.

Because the data available for our application of interest (*target data*) are limited, developing an accurate model using only these data is infeasible. Instead, we investigate transfer learning approaches to utilize data from other cardiac surgery procedures as well as from other institutions (*source data*). We first evaluate the effectiveness of leveraging information across procedures within a single hospital. We achieve significant improvements over baseline: at Hospital 1, the average AUC for operative mortality increased from 0.58 to 0.70. However, not all source examples are equally useful.

Next, we evaluate the effectiveness of leveraging data across hospitals. We show that leveraging information across hospitals has variable utility; although it can result in worse performance (average AUC for stroke at Hospital 1 dropped from 0.61 to 0.56), it can also lead to significant improvements (average AUC for operative mortality at Hospital 1 increased from 0.70 to 0.72).

Finally, we present an automated approach to leveraging the available source data. We investigate how removing source data based on how far they are from the mean of the target data affects performance. We propose an instance-weighting scheme based on these distances. This automated instance-weighting approach can achieve small, but significant improvements over using all of the data without weights (average AUC for operative mortality at Hospital 1 increased from 0.72 to 0.73). Research on these methods can have an important impact on the development of clinical risk-stratification tools targeted towards specific patient populations.

Thesis Supervisor: John V. Guttag
Title: Dugald C. Jackson Professor

Thesis Errata Sheet

Author Jen Gong

Primary Dept. Electrical Engineering and Computer Science

Degree M.S. **Graduation date** 06/2014

Thesis title

Improving Clinical Risk-Stratification Tools: Instance-Transfer for Selecting Relevant Training Data

Brief description of errata sheet

On page 50: Misstated the normalization method that was used.

Number of pages 1 (11 maximum, including this page)

▶ **Author**: I request that the attached errata sheet be added to my thesis. I have attached two copies prepared as prescribed by the current *Specifications for Thesis Preparation*.

Signature of author Signature redacted Date 3/5/15

▶ **Thesis Supervisor** or **Dept. Chair**: I approve the attached errata sheet and recommend its addition to the student's thesis.

Signature Signature redacted Date 3/5/15

Name John Guttag Thesis supervisor Dept. Chair

▶ **Dean for Graduate Education**: I approve the attached errata sheet and direct the Institute Archives to insert it into all copies of the student's thesis held by the MIT Libraries, both print and electronic.

Signature Signature redacted Date 3/20/15

Name Christine Ortiz

1. **(a) Page 50, Lines 18-20.**

   (b) Current text: "Missing data for continuous features were replaced with the mean, and features were normalized to have a range between 0 and 1 by subtracting the minimum of the feature and dividing by the range."

   (c) Correct text: "Missing data for continuous features were replaced with the mean. Continuous features were also normalized by subtracting the 1st percentile of the feature values and dividing by the difference between the 99th and the 1st percentiles."

Acknowledgments

For the last two years, I have been lucky to have been part of a fantastic research group and to have had John Guttag as my thesis advisor. I have received invaluable guidance from John, and he has pushed me to think deeply about the technical aspects of my research as well as how my work can have practical impact. John's encouragement, optimism, and enthusiasm have made our research efforts more rewarding.

I would like to thank Collin Stultz and Zeeshan Syed for our many productive discussions. I would also like to thank our clinical collaborators for their medical insight.

I am grateful to my labmates for their technical guidance and their moral support. Jenna, Garthee, Anima, Yun, Guha, Amy, and Joel have provided me with enriching technical discussions as well as a wonderful lab ethos.

Finally, I would like to thank my family. Their love and support mean the world to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Risk-stratification models are increasingly important in the clinical decision-making process. Often, one of the challenges to developing accurate models is the limited amount of relevant data. Many real-world applications involve small, incomplete, and noisy data sets with high class-imbalance. On top of this, such models are often developed for very specific patient populations; many exclusion criteria are applied to the *available* data to obtain the *relevant* data. The combination of these effects means there may not be enough relevant data to develop and test a model.

In our application of interest, we start off with a dataset consisting of over 20,000 surgeries, from two hospitals (Hospital 1 and Hospital 2). We are interested in developing models for risk-stratifying patients who receive an isolated aortic valve replacement (i.e., an aortic valve replacement (AVR) without any other procedure) for an adverse outcome (e.g., operative mortality, stroke) at a given hospital. Already, we have applied a number of exclusion criteria. If we consider only surgeries at Hospital 1, we reduce our dataset from over 20,000 to approximately 10,000. If we consider only isolated AVRs at Hospital 1, we reduce that number even further, to 917. These 917 surgeries contain only 19 cases of operative mortality and 21 strokes. If we want to train and test a model on this small set, we have to divide it even further.

Fortunately, there is useful information from the thousands of other kinds of cardiac surgeries that can be taken advantage of. Using these data has the potential to significantly improve the performance of our model. This is an example of *transfer* *learning*.

When using traditional machine learning methods, one assumes that the training and test data lie in the same feature space and are drawn from the same distribution. In some cases, such as ours, there are not enough training data available for the task of interest (the *target task*), but there are related *source* (auxiliary) data available. The data pertaining to the target task are said to lie in the *target domain*, and the source data to lie in the *source domain*. Here, the target task is risk-stratifying patients receiving isolated AVR, and the source data are other cardiac surgeries.

For our application, we consider the subfield of *inductive transfer learning*, where labels are available for both the target and source data. There are several methods of transferring information in the inductive setting: instance-transfer, feature-representation transfer, and parameter-transfer [1]. Our work focuses on instance-transfer. We will investigate what subsets of source data, when added to target task training data, result in the best performance. Although in our application, the target and source data lie in the same feature space, they are not drawn from the same distribution. Augmenting our target task training data with source data is therefore not guaranteed to improve performance on the target task; the source data could have no effect, or they could even hurt performance. Instance-transfer techniques seek to make use of the helpful instances in the source data while discounting the harmful ones.

1.1 Problem Formulation

We have data collected for the STS National Adult Cardiac Surgery Database from two hospitals in the United States. These data were collected from 2001 to 2011. We consider the preoperative data from patients who were placed on cardiopulmonary bypass for the full duration of the cardiac surgery. Our data include patient risk factors, demographics, the hospital where the patient received the procedure, and operative variables identifying the procedure(s) a patient received (e.g., an aortic valve replacement (AVR), a coronary artery bypass grafting (CABG), an AVR and a

CABG together, etc.). We represent these data as $D = \{(\mathbf{x}, y)\}$, where each feature vector $\mathbf{x} = [x_1, x_2, ..., x_d]$ lies in the same feature space $\mathcal{X}$ with dimensionality d, and each corresponding outcome y lies in the space $\mathcal{Y} = \{-1, +1\}$. $y = +1$ if the adverse outcome of interest (e.g., stroke) occurs and $y = -1$ otherwise.

In this context, a *target task* can be defined by specifying the following:

1. a patient population of interest, where the selection criteria are fulfilled if the value of the characteristic function $f : \mathcal{X} \rightarrow \{0, 1\}$ for a patient with feature vector $\mathbf{x} \in \mathcal{X}$ is 1, and

2. an adverse outcome y for each patient in this population.

Then the target and source data can be defined as follows:

- $T = \{(\mathbf{x}, y) | f(\mathbf{x}) = 1\}$

- $S = \{(\mathbf{x}, y) | f(\mathbf{x}) = 0\}$

Although many different patient selection criteria could be of interest, in this work, we examine characteristic functions based only on which hospital the procedure was performed at and whether or not the procedure performed was an isolated aortic valve replacement. We chose to allow for separating the data by hospital because it has been established in the literature that accounting for hospital-specific characteristics is important in developing risk-stratification tools [2, 3]. We chose isolated AVR as a target task because because there has been an increased focus on the development and validation of risk models for isolated AVR. This was stimulated by the relatively recent introduction of an alternative intervention for high-risk patients with severe, symptomatic aortic stenosis who cannot receive surgical AVR [4, 5, 6, 7, 8], called transcatheter aortic valve replacement (TAVR). Risk models such as the STS risk score [9, 10] and the logistic EuroSCORE [10, 11] have been used to determine who is eligible to receive this treatment.

We thus have two target tasks. For task h ($h = 1, 2$) the target data T_h are chosen according to $f_h(\mathbf{x})$. $f_h(\mathbf{x})$ is 1 if the patient received a surgery at hospital h and that surgery was an isolated AVR, and 0 otherwise. The example is included in the source data S_h if $f_h(\mathbf{x}) = 0$.

In the following chapters, we will explore how augmenting the target task training data with source data can improve performance on the target task. We hypothesize that for each task h, augmenting the target task training data with a subset of source data similar to the target data will improve performance on the target task more than augmenting with all of the available source data S_h. More formally,

- Consider choosing $s \subset S_h$ such that the probability that an example in s is drawn from the empirical joint distribution of the target data $\hat{p}_{T_h}(\mathbf{x}')$ is greater than the probability that an example in $S_h - s$ is drawn from $\hat{p}_{T_h}(\mathbf{x}')$, where $\mathbf{x}'$ contains a subset of the features in the full feature vector $\mathbf{x}$.

- We hypothesize that augmenting target task training data with s will improve performance on the target task more so than augmenting it with S_h.

To evaluate how likely it is for an example in s to be drawn from $\hat{p}_{T_h}(\mathbf{x}')$, we consider the distance of the source data from the target data. The closer examples in the source data are to the target data, the more similar they are. We compute these distances using different subsets of the features in $\mathcal{X}$. We map the source and target data from $\mathcal{X} \rightarrow \mathcal{X}'$, where $\mathcal{X}'$ includes only the features in the data in the chosen subset. For all of the following experiments, we will use this subspace $\mathcal{X}'$ to evaluate the distance of each source example from the target data. However, we will then use the full set of features $\mathcal{X}$ to train and evaluate the model.

1.2 Related Work

1.2.1 Transfer Learning

Identifying auxiliary data that are likely to be useful for the target task is an important question in the area of transfer learning. Although previous work has shown that the use of source data can result in significantly improved performance on the target task, how to correctly ascertain which data are useful is still an open question.

Transfer learning assumes that the joint distribution of the features and outcome of the source data are related to the joint distribution of the features and outcome of the target data. When this assumption is not satisfied, it can result in negative transfer [12]. However, evaluating whether or not this assumption holds is non-trivial. Most methods that seek to address this problem try to manipulate the source joint distribution to be closer to the target joint distribution. These methods often focus on moving either the marginal distribution of the source data, $p_s(\mathbf{x})$, closer to that of the target data, $p_t(\mathbf{x})$, or the conditional distribution $p_s(y|\mathbf{x})$ towards $p_t(y|\mathbf{x})$. Some work has been done on shifting both the marginal and conditional closer to the target distribution [13]. These *domain adaptation* methods take two approaches: 1) instance-weighting or instance-selection [14, 15], and 2) changing the feature space [13, 16, 17, 18]. Instance-weighting approaches focus on minimizing the distance between the distributions of the source and target data by giving lower weight or removing instances from the source data which appear to be very different from the target data. Feature-representation transfer approaches, on the other hand, focus on finding a feature space in which the distance between the source and target distributions is minimized. We will focus on the instance-weighting approaches here.

These approaches rely on having a similarity or distance metric that accurately reflects how far the source distribution is from the target distribution. In [15], the authors consider a natural language processing task, where there are labeled data available in the source domain, but little or no labeled data in the target domain. The authors use a heuristic method to prune instances based on how different $p_t(y|x)$ and $p_s(y|x)$ are. They identify the need for instance adaptation based on the marginal distributions, but do not explore a solution to this problem in their experiments. The authors of [14] propose a method called *Transfer AdaBoost*, which assumes that the source data, while drawn from a different distribution than the target data, may still have useful information. They augment target task training data with these source data and decrease the weights of source instances which are misclassified at each iteration. Their hypothesis is that the misclassified source data are the most dissimilar from the target data. These methods require training models in order to determine which instances should be pruned or used with lower weight.

In contrast, [19] uses target data to identify similar instances in the auxiliary data. They use each target task training example to extract similar examples from the auxiliary data. They explore two approaches to using the auxiliary data. The authors incorporate a term in the objective function for fitting the auxiliary data, where a parameter γ controls how important fitting the auxiliary data is. They explore how the auxiliary data can be incorporated into the framework of kNN and SVMs. They demonstrate that including the auxiliary examples when learning a model can significantly improve accuracy on the target task.

1.2.2 Cardiac Surgery Risk Models

In the field of cardiac surgery, many risk-stratification models have been developed for adverse outcomes such as death and stroke. These risk models range from general cardiac surgery risk stratification tools, such as the additive and logistic EuroSCOREs [20, 21], to models developed for specific cardiac surgeries at specific institutions.

Models developed for specific cardiac surgeries at specific institutions have much less available data than more global models; however, they are able to account for hospital-specific information. Global models allow for the use of much more data, either from multiple hospitals or encompassing multiple surgeries, but they do not account for institution-specific or surgery-specific differences.

The national risk models for adult cardiac surgery developed and published by the Society for Thoracic Surgeons (STS) are for several distinct categories of surgeries: 1) coronary artery bypass grafting (CABG) [22], 2) isolated valve surgery [23], and 3) valve plus CABG surgery [24]. The EuroSCORE models, on the other hand, were developed and tested on all patients receiving cardiac surgery while on cardiopulmonary bypass [20, 21, 25]. In the context of surgical procedures, the STS models can be viewed as *local* models, while the EuroSCORE models are *global* models. Although the framework of the global model incorporates more data, it may perform poorly when applied to a specific subset of the patients (e.g., patients receiving one particular surgery).

The additive and logistic EuroSCOREs have been shown to overestimate the risk associated with aortic valve replacements [7, 26, 5]. The authors of [7] cite the fact that the EuroSCORE was developed on a population where most of the patients received surgery for coronary artery disease, rather than isolated AVR, as one possible reason for why the additive and logistic EuroSCORE do not perform well on isolated AVR. On the other hand, while the local models may be able to better learn the specific characteristics of patients receiving a particular type of surgery, each model has much less data available for development and testing.

Finally, both the STS and EuroSCORE models are global in the context of institutions; all of the models were developed on multi-center datasets. Institutions vary in size, location, patient population, staff, etc. Because of such differences, a general model learned on all accumulated data may not perform well on each individual hospital. In fact, previous work on risk-stratification for other outcomes has shown that accounting for variations between hospitals and using hospital-specific features can improve performance [2, 3]. In the field of cardiac surgery risk models, the authors of [27] demonstrated that a "ready-made" risk model (one developed and tested on data external to the institution of interest) did not perform as well on their study population as a model that considered the same risk factors as the "ready-made" model but was recalibrated to their specific institution. This "ready-made" model also did not perform as well as a new model that incorporated other risk factors. No correction for institutional differences has been published in the area of cardiac surgery, but models such as the STS and EuroSCORE models are still used in hospitals to estimate a patient's preoperative risk of mortality and postoperative complications.

Previous work has been published with validation of these global models on single institutions [28, 29, 30, 31, 32, 33]. Many of them report good discrimination performance, but less reliable calibration performance. When the additive and logistic EuroSCOREs were validated on individual institutions, they overestimated the risk of mortality [32, 34, 28]. The EuroSCORE II, launched in 2011, updated the EuroSCORE risk model on a contemporary cohort of patients; a logistic model was developed and tested on this population [25]. However, [35] found that the EuroSCORE

II model did not perform better than other risk models (including the original additive and logistic EuroSCOREs) on the task of predicting mortality in patients who received either surgical AVR or transcatheter AVR.

Thus, there is room for improvement in present cardiac surgery risk models. We believe that by viewing different institutions as separate but related tasks (as in [2, 3]) and different *surgical procedures* as separate but related tasks, we can build a model that is able to take advantage of all of the available data while also accounting for institution-specific and procedure-specific characteristics.

1.3 Contributions

The contributions of this thesis are as follows:

- We demonstrate that using only the target data (isolated AVR at hospital h) to develop and test a model results in unsatisfactory discrimination performance.
- We show that augmenting target task training data with subsets of source data from the same hospital improves performance. We demonstrate that the degree of improvement depends on 1) the size of the subset, and 2) how similar the subset is to the target data.
- We show that augmenting target task training data with subsets of source data from *both* hospitals can result in *negative transfer*. We demonstrate that using all of the available data is not always the optimal choice.
- We propose an automated method by which a subset of source data can be chosen from all of the available source data by considering the distance of each example from the target task training data. We show that considering only some portion of the available source data can result in better performance than using all of the available data in the context of this automatic approach. Previous work has either pooled many different types of surgeries together [20, 21, 25, 36], or segmented them based on expert knowledge [22, 23, 24].

- Lastly, we show that instead of augmenting the training data with a subset of source data, using all of the available data in combination with an instance-weighting scheme can result in the best or close to the best performance out of all of the approaches. Additionally, this final method requires no user-specified parameters; it is completely automated.

1.4 Outline

In Chapter 2, we describe the data we use as well as the application we consider in more detail. We outline our data processing and model development and validation methods.

In Chapters 3 and 4, we consider subsets of S_h according to how similar they are to T_h based on two criteria: 1) whether or not an AVR was performed, and 2) whether or not the procedure was performed at hospital h. In Chapter 3, we focus on the case where for each target task h, we only consider the subset of S_h from the same hospital. Then the available source data vary only in whether or not an AVR was performed.

In Chapter 4, we first consider the subset of source data where the procedure is the same as in the target data (isolated AVR). For a target task h, these data must come from the other hospital. Lastly, we explore how augmenting the target training data with subsets of source data from the other hospital affects performance. As in Chapter 3, we consider subsets of source data that vary in in whether or not an AVR was performed. This final method transfers information across procedures and across hospitals.

In Chapter 5, we use the full feature set $\mathcal{X}$ to evaluate similarity of source examples to target examples. We investigate automated methods to choosing the best subset of source data to augment our target task training data with. We compare choosing the best subset with *instance-weighting*, where all source examples are used but are assigned different weights based on some measure of their estimated utility.

Finally, in Chapter 6, we briefly summarize the contributions of this thesis and discuss future work.

Chapter 2

Data and Background

In this chapter, we first provide some background on the data we use and the application of risk-stratifying patients receiving isolated AVR. Next, we discuss preprocessing of the data. We then compare the demographics, risk factors, operative features, and post-operative outcomes of patients in the national database and patients from the two hospitals in our data. Because each STS model is developed for patients receiving a specific set of procedures, we focus on the data from our application of interest (isolated aortic valve replacements). We show that not only are there differences between the hospitals and the national database, there are also differences between the two institutions. Finally, we describe our model development and validation methods.

2.1 Aortic Valve Replacements

We consider the application of developing risk stratification models for adverse outcomes resulting from an isolated aortic valve replacement. AVR is the standard of care for treatment of severe, symptomatic aortic stenosis. Aortic stenosis is a condition typically caused by calcification of the aortic valve leaflets [37]. The aortic valve is the gateway between the left ventricle and the aorta. When the valve leaflets become calcified, the effective valve opening is restricted. Thus, the valve will not open fully, and in some cases, it will not close completely either. This change is shown in Figure 2-1. The left ventricle must work harder to compensate for this reduced valve opening. The changes in the left ventricle occur gradually; patients may be asymptomatic even as the left ventricle struggles to compensate for the stenotic valve [37]. The common symptoms that physicians look for are shortness of breath, dizziness, and angina after exercise. Eventually, untreated aortic stenosis will lead to heart failure, syncope, angina (chest pain), and death [37].

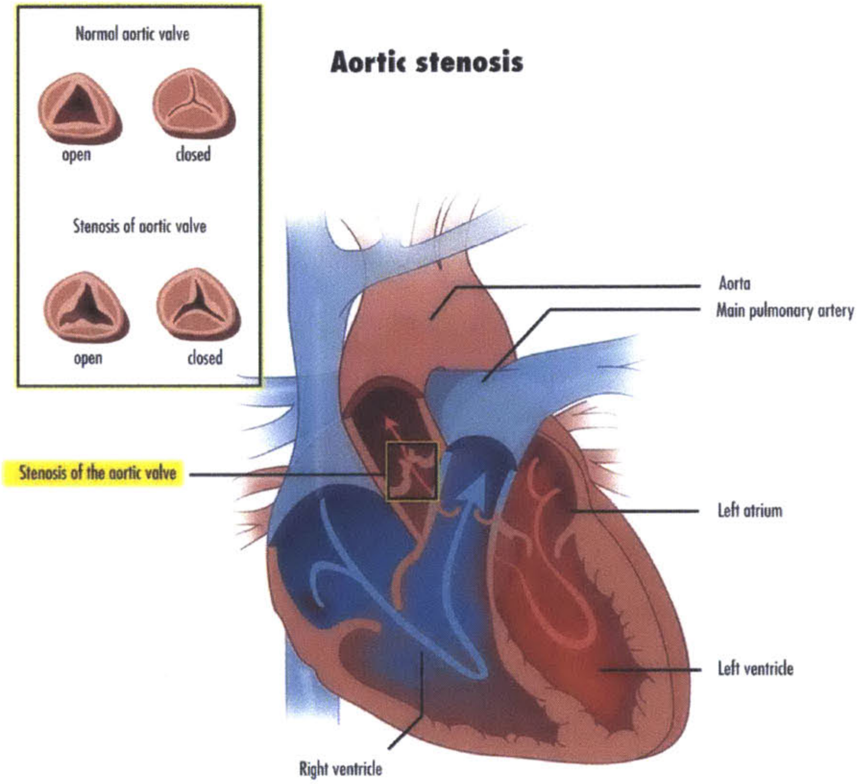

Figure 2-1: A comparison of a normal aortic valve and a stenotic valve. [1]

After individuals with aortic stenosis begin exhibiting symptoms, they have on average two to three years to live [37] without treatment. Surgical aortic valve replacement, an open-chest surgery that typically requires the use of cardiopulmonary bypass, often results in increased quality of life and longer life-expectancy [37]. However, patients who are determined to be too high-risk may not be able to receive this surgery.

[1] This image was taken from http://www.childrenshospital.org/health-topics/conditions/aortic-valve-stenosis.

An alternative treatment called a transcatheter aortic valve replacement (TAVR), first successfully performed on a human in 2002 [38], was approved by the FDA in the United States for use in patients who cannot receive surgical AVR in 2011 [39]. In 2013, the FDA expanded the use of this treatment to allow patients deemed at high-risk for surgical AVR to receive it [40]. This treatment uses a transcatheter approach to insert a new valve. It does not require the use of cardiopulmonary bypass and it is minimally invasive. However, in the Placement of Aortic Transcatheter Valves (PARTNER) trial, a randomized, multi center clinical trial, more major strokes were observed in the patients that received TAVR as opposed to standard AVR [9]. Additionally, transcathether AVR was associated with a higher incidence of major vascular complications within 30 days [9]. Despite these complications, for high-risk patients, transcatheter AVR presents a possible alternative treatment to surgical AVR.

Accurately determining a patient's risk of mortality or other severe complications of surgical AVR has proven difficult using existing clinical risk models [41, 42, 7, 43, 26, 44]. Improvement of these methods is necessary for an accurate cost-benefit analysis of the possible treatments for each individual patient.

2.2 Data

2.2.1 Society of Thoracic Surgeons National Adult Cardiac Surgery Database

The Society of Thoracic Surgeons (STS) National Adult Cardiac Surgery Database was established in 1989. It contains data from over 90% of the institutions in the United States that perform cardiac surgery [45]. Its primary purpose is to develop risk models that allow for accurate quality assessment and provider comparison by controlling for case mix differences between patient populations at different hospitals [22, 23, 24]. However, these models have also been used to provide estimates of patients' preoperative risk for adverse outcomes such as mortality and stroke [9].

2.2.2 Consolidating different data versions

Our data were collected using three different data version guidelines in the STS database, versions 2.41, 2.52, and 2.61. With each change in the guidelines, there were changes in the variables in the database, changes in the names of the features, and changes in definitions of features with the same name. As an example, in version 2.41, the variable "Arrhythmia" encoded whether or not there was an arrhythmia present within two weeks of the procedure. If there was, the variable "Arrhythmia Type" encoded whether the arrhythmia was "Sustained Ventricular Tachycardia or Ventricular Fibrillation requiring cardioversion and/or IV amiodarone," "Heart block," or "Atrial fibrillation/flutter requiring Rx." In version 2.52, the definition of "Arrhythmia" changed: it encoded "whether there is a history of preoperative arrhythmia," but did not indicate whether or not that arrhythmia was present within two weeks of the procedure. However, "Arrhythmia Type" still encoded which arrhythmia was present within two weeks of the procedure, with the additional option of "None." Finally, in 2.61, the categorial variable "Arrhythmia Type" was removed and was replaced by binary indicator variables for each option (sustained ventricular tachycardia or fibrillation, heart block, and atrial fibrillation or flutter).

Where it was possible, we created mappings from semantically equivalent features in each data version to consistent binary indicator features. Most of these mappings were between version 2.61 and the other versions. Because the national STS model was developed and tested entirely on data from versions 2.41 and 2.52, it did not have to deal with the majority of the changes. To avoid removing all data from version 2.61 (approximately a third of the available data), we instead preserved as many features as possible and discarded features that were unique to a single data version and had no clear mapping to features in the other data versions.

The specifications for data versions 2.41, 2.52, and 2.61 were retrieved online on the STS website [46, 47, 48].

2.3 Comparing the STS National Database to our data

We compared the proportion of patients in each hospital to the proportion of patients in the national database who had each risk factor. These risk factors were found in [23].

To test for statistical significance, we used a z-test to compare the proportions. The z-test uses the test statistic given in Equation 2.3:

$$p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} \tag{2.1}$$

$$SE = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.2}$$

$$z = \frac{p_1 - p_2}{SE} \tag{2.3}$$

where p_1 and p_2 are the proportions in populations 1 and 2, respectively, and n_1 and n_2 are the sizes of those populations. We then compute the two-tailed p-value by using the standard normal cumulative distribution function to compute the probability of observing a value less than $-|z|$ or greater than $|z|$.

2.3.1 Demographics

This section details the demographic differences in the populations of the two institutions and the national study population. We used the information provided in [23] about the distribution of features in the national study population. The results are shown in Table 2.1; differences from the national population with p-value < 0.01 are in bold.

We note that there are significant differences in the demographics of the population. In particular, the proportion of patients with age < 55 is much lower in the two hospitals in our data than in the national study population. Additionally, there is a greater proportion of older patients (≥ 75) in the two institutions we consider

Table 2.1: Demographics

	2008 National Study Population ($n = 67,292$)	Hospital 1 ($n = 917$)	Hospital 2 ($n = 1602$)
Age $<$ 55	19.66	**14.61**	**15.98**
Age 55-64	19.39	17.12	17.54
Age 65-74	27.19	28.14	25.28
Age $\geq$ 75	33.85	**40.13**	**41.20**
Male	58.27	57.91	55.49
Female	41.73	42.09	44.51
Caucasian	87.17	**94.11**	**95.51**
Black	5.28	**1.09**	**1.31**
Hispanic	3.48	**1.74**	**2.00**
Asian	1.07	0.76	0.50

than in the national study population. This suggests that our study population is a higher-risk population.

2.3.2 Postoperative outcomes

In this thesis, we will focus on two post-operative outcomes: operative mortality and stroke. These are defined in version 2.61 of the STS database as follows:

Operative Mortality: "Includes both (1) all deaths occurring during the hospitalization in which the operation was performed, even if after 30 days; and (2) those deaths occurring after discharge from the hospital, but within 30 days of the procedure unless the cause of death is clearly unrelated to the operation."

Stroke: "Whether the patient has a postoperative stroke (i.e., any confirmed neurological deficit of abrupt onset caused by a disturbance in cerebral blood supply) that did not resolve within 24 hours." Complications such as stroke are recorded through "the entire postoperative period up to discharge, even if over 30 days." [48].

We compared the proportions of adverse outcomes occurring in the national database (found in [23]) with those in the two hospitals. These results are shown in Table 2.3.2. Hospital 2 has a significantly different proportion of strokes as com-

Table 2.2: Percentage of adverse outcomes occurring in the 2008 national study population, Hospital 1, and Hospital 2 (isolated AVR only).

	2008 National Study Population ($n = 67,292$)	Hospital 1 ($n = 917$)	Hospital 2 ($n = 1602$)
Operative Mortality	3.21	2.07	2.81
Stroke	1.50	2.29	**2.87**

pared with the national database ($p < 0.001$).

2.3.3 Risk Factors

We excluded features that did not have corresponding mappings in version 2.61 and comparisons between proportions of missing data. To provide a concise picture of the differences between the national study population and the two hospitals in our data, for risk factors which had only 3 possible unique values (present, not present, or missing), we only compared the proportions of patients where the risk factor was present. Out of 103 comparisons of pre-operative risk factors between the national database and the two hospitals in our data, over half of the features demonstrated differences at a significance level of 0.01 with at least one of the hospitals in our data. The features that demonstrate significant differences (p-value < 0.01) between the national study population and either of the two hospitals in our data are shown in Table 2.3. Features that were also significantly different between the two hospitals (p-value < 0.01) are in bold.

Table 2.3: Differences between National Database and Hospitals 1 and 2.

	2008 National Study Population	**Hospital 1**	**Hospital 2**
	($n = 67,292$)	($n = 917$)	($n = 1602$)
Body surface area, 1.50-1.74 m^2	20.38	22.25	23.78
Body surface area, > 2.00 m^2	38.65	36.97	34.58
Body mass index, < 25 kg/m^2	27.51	30.97	30.90
Body mass index, kg/$m^2 \geq 35$	14.87	14.72	11.74
No diabetes	77.35	77.64	80.52
Hypertension	**66.60**	**72.74**	**66.04**
Hypercholesterolemia	50.33	69.68	65.11
No chronic lung disease	79.51	83.97	84.14
Mild chronic lung disease	**10.39**	**10.03**	**14.73**
Moderate chronic lung disease	**5.98**	**3.27**	**0.69**
Severe chronic lung disease	**3.14**	**2.73**	**0.37**
No CVA	92.91	93.68	95.88
Remote CVA (> 2 weeks)	6.25	5.67	3.81
No endocarditis	94.00	93.46	95.57
Active endocarditis	3.07	3.16	1.75
Creatinine < 1.00 mg/dL	**38.16**	**30.21**	**47.82**
Creatinine 1-1.49 mg/dL	**47.64**	**53.44**	**43.26**
Creatinine 1.50 - 1.99 mg/dL	**7.55**	**11.12**	**6.49**
Creatinine ≥ 2.50 mg/dL	1.11	2.62	1.37
Dialysis	**2.18**	**1.96**	**0.62**
Immunosuppressive treatment	3.08	7.20	4.99
Previous valve surgery	6.22	5.56	7.80
No PCI	92.35	90.73	89.89
PCI within 6 hours	0.09	0.00	0.69
PCI not within 6 hours	6.95	9.16	9.36
Elective status	**76.88**	**70.67**	**83.83**
Urgent status	**21.80**	**28.90**	**15.42**
No prior MI	**90.43**	**86.91**	**90.45**
MI 8-21 days	**0.71**	**3.05**	**0.62**
No angina	**73.67**	**63.58**	**84.33**
No arrhythmia	**85.38**	**79.39**	**90.95**
AFib/flutter	**11.25**	**16.58**	**6.49**
Heart block	**1.65**	**5.56**	**1.12**
Sustained VT/VF	**0.72**	**1.09**	**0.12**
NYHA Class I	**15.19**	**0.11**	**1.50**
NYHA Class II	**30.16**	**3.05**	**11.36**
NYHA Class III	**37.87**	**11.23**	**19.91**
NYHA Class IV	**12.04**	**10.91**	**3.12**
Congestive heart failure	**37.43**	**25.30**	**35.89**
No diseased coronary vessels	81.84	75.25	76.15
One diseased coronary vessel	8.01	11.78	10.80
Two diseased coronary vessels	3.24	4.69	5.56
Three diseased coronary vessels	5.60	8.29	7.49
Left main disease, ≥ 50%	1.67	1.96	3.43
Ejection Fraction, < 25 %	**2.64**	**2.62**	**0.69**
Ejection Fraction, 25-34 %	5.66	4.91	3.00
Ejection Fraction, 35-44 %	9.19	5.02	5.74
Ejection Fraction, 45-54 %	18.44	9.05	10.42
Ejection fraction, ≥ 55%	54.37	78.19	76.90
Aortic stenosis	**79.83**	**88.55**	**82.77**
Mitral stenosis	2.08	5.89	3.81
No aortic insufficiency	**38.43**	**24.21**	**37.33**
Trivial aortic insufficiency	**8.79**	**19.74**	**12.23**
Mild aortic insufficiency	**14.88**	**29.88**	**24.03**
Severe aortic insufficiency	23.08	13.20	12.17
No mitral insufficiency	**60.12**	**13.63**	**21.54**
Trivial mitral insufficiency	**10.83**	**27.48**	**22.53**
Mild mitral insufficiency	19.42	39.26	40.01
Moderate mitral insufficiency	6.6	17.23	13.73
No tricuspid insufficiency	**74.27**	**21.92**	**35.71**
Trivial tricuspid insufficiency	**8.34**	**35.01**	**28.34**
Mild tricuspid insufficiency	10.9	29.66	25.66
Moderate tricuspid insufficiency	3.16	9.60	8.55
Severe tricuspid insufficiency	0.44	1.31	0.56
No pulmonic insufficiency	**89.85**	**45.37**	**73.91**
Trivial pulmonic insufficiency	**3.52**	**39.15**	**17.35**
Mild pulmonic insufficiency	1.99	7.74	6.49
Moderate pulmonic insufficiency	0.31	1.09	0.75

2.4 Model development and validation methods

Although we used data collected for the STS database, we did not utilize any of the expert-driven features included in the STS risk models. These included modeling some continuous features using linear splines with specified knots, as well as interaction terms between features [22, 23, 24].

2.4.1 Feature construction

Binary features were left unchanged. Categorical features were replaced with binary features for each unique value of that feature. Finally continuous features were discretized by quintile, according to the training data, and the resulting categorical features were replaced with binary features. The final feature vectors consisted of 251 binary features.

2.4.2 Model development

We learned all of our models using LIBLINEAR [49]. All models are produced using L2-regularized logistic regression with an asymmetric cost parameter [50], as given by Equation 2.4, where C_+ and C_- are the costs of misclassifying examples in the positive class and the negative class, respectively.

$$\begin{aligned}\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+ &\sum_{i:y_i=+1} \log(1+e^{-y_i\mathbf{w}^T\mathbf{x}_i}) \\ + C_- &\sum_{i:y_i=-1} \log(1+e^{-y_i\mathbf{w}^T\mathbf{x}_i}) \qquad (2.4)\end{aligned}$$

Logistic regression outputs a continuous measure that, for our application, represents the risk of the patient for the adverse outcome of interest. We used L2-regularization to help prevent overfitting.

We divided the target data into sets of 50% training and 50% test, stratified by outcome. This was repeated 100 times. All of our methods were applied to these same 100 test sets and incorporated the corresponding target training data into the training set. We used 5-fold cross-validation on each training set to determine the hyper-parameters of the model. We optimized for the area under the receiver operating characteristic curve (AUC). We considered two hyperparameters: the cost for misclassifying an example in the negative class, C_-, and the cost of misclassifying an example in the positive class, C_+. These parameters are typically found by performing a grid search over C_- and the ratio $\frac{C_+}{C_-}$ (the asymmetric cost parameter). To speed up computation, we set the value of $\frac{C_+}{C_-}$ to a constant equal to the class imbalance in the training data $(\frac{\sum_{i=1}^{n} 1(y_i=+1)}{\sum_{i=1}^{n} 1(y_i=-1)})$. We searched the space from 10^{-7} to 10^{2} in powers of 10 for C_+ and used cross-validation to select the best value.

This process is diagrammed in Figure 2-2.

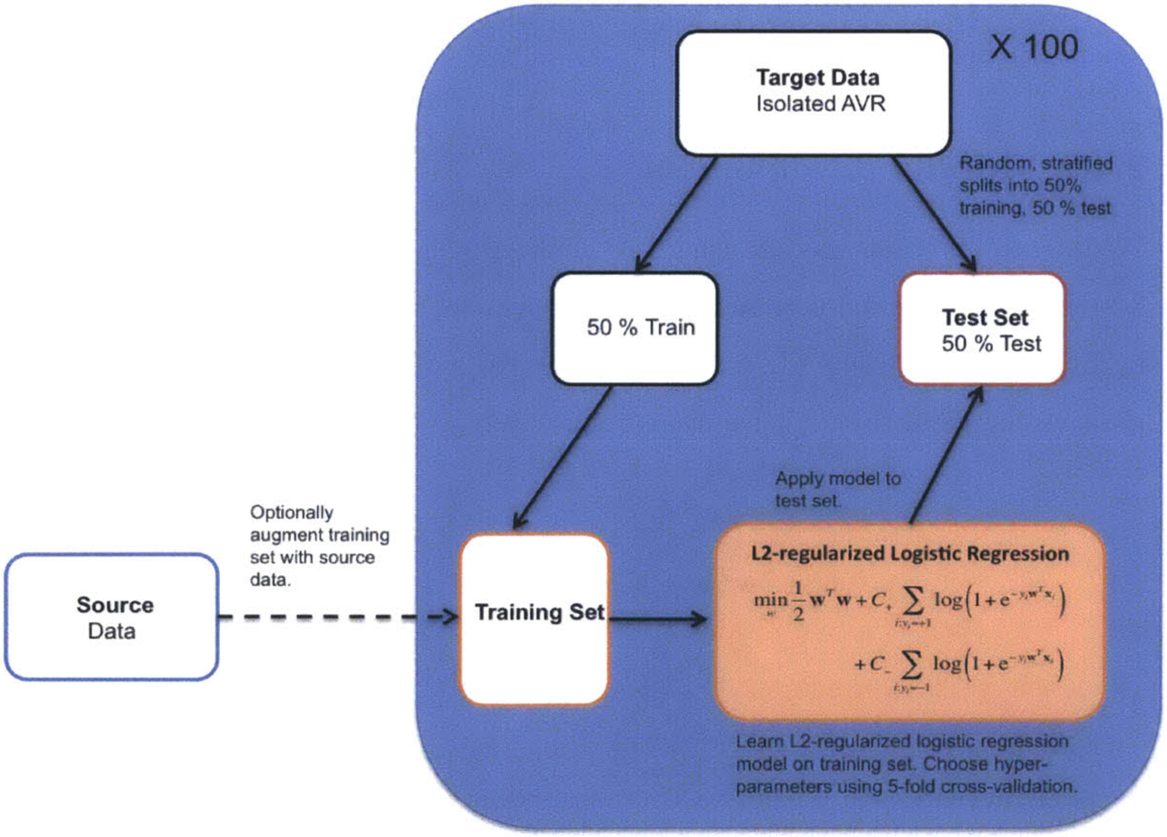
X 100
Target Data
Isolated AVR
Random, stratified splits into 50% training, 50 % test
50 % Train
Test Set
50 % Test
Apply model to test set.
Optionally augment training set with source data.
Source Data
Training Set
L2-regularized Logistic Regression
$\min_{w} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+ \sum_{i:y_i=+1} \log\left(1+e^{-y_i\mathbf{w}^T\mathbf{x}_i}\right) + C_- \sum_{i:y_i=-1} \log\left(1+e^{-y_i\mathbf{w}^T\mathbf{x}_i}\right)$
Learn L2-regularized logistic regression model on training set. Choose hyper-parameters using 5-fold cross-validation.

Figure 2-2: Model development procedure.

2.4.3 Evaluation

We evaluate each of our methods using the area under the receiver operating characteristic curve (AUC), the precision, and the recall. We report each value along with its 95% confidence interval (CI).

AUC

The ROC curve depicts the tradeoff between the false-positive rate and the true-positive rate. The AUC is used to summarize the ROC curve. An AUC of 0.5 means a classifier does no better than randomly assigning the outcomes; it has no discriminative value. A classifier with an AUC of 1 has perfect discrimination; it is able to perfectly distinguish the adverse events from the non-adverse events. For most classifiers, the AUC will lie between these two values. Generally, a value of 0.7 is used as a standard for good performance.

Precision and Recall

Precision and recall are evaluated at one point along the ROC curve. We considered the upper quintile of risk as high-risk; that is, instances where the patient was assigned a probability above the 80th percentile were considered high-risk (+1), and individuals below that threshold were considered low-risk (-1). Precision and recall are defined as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}},$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

To give the reader an idea of how our methods perform, we describe the performance of a random classifier in terms of precision and recall here. While an AUC of 0.5 always indicates the classifier does no better than randomly assigning the outcomes, precision depends on the incidence of the outcome in the population, and recall depends on the definition of high-risk. We call the incidence probability p. Because we consider the upper quintile of risk as high-risk, 20% of the test cases, regardless of the classifier, will be classified as positive. Thus, the precision of a random classifier is p, and the recall is 0.20. This provides a baseline against which other classifiers can be compared.

Comparing Models

To determine whether the difference in performance between two models was statistically significant, we used a paired t-test [51] and a paired Wilcoxon signed-rank test [52]. A paired t-test is a parametric test used to measure the statistical significance of paired differences (e.g., when comparing a group of patients' heart rates before and after exercise to determine if exercise increases heart rate). The Wilcoxon signed-rank test is the non-parametric equivalent. The t-test makes assumptions of normality, whereas the Wilcoxon signed-rank test does not. For completeness, we use both tests to evaluate significance; we only consider results where both p-values are less than 0.05 to be significant.

In each test, the null hypothesis is that there is no difference between the performance of the two classifiers. This is done by comparing the average difference between the performance of the two classifiers on the 100 test sets with 0. For all statistical significance tests, we computed the two-tailed p-value. This two-tailed p-value, when below a specified threshold (e.g., 0.05, 0.01, etc.), signifies that one classifier performs significantly better than the other.

2.5 Summary

In this chapter, we showed that the patient populations of the two institutions we consider demonstrate significant differences from the national study population from which the 2008 STS risk models were developed and validated. These differences are evident in the demographics, the risk factors of the patients, and the outcomes. Inter-institutional differences are also evident in the proportions of risk factors. We also outlined our methods for processing the data as well as for model development and validation.

Chapter 3

Leveraging information across cardiac procedures at a single institution

In this chapter, we consider transfer of knowledge within a single institution.

We consider two ways in which subsets of source data can vary: 1) size, and 2) similarity to the target data. For each target task $h = 1, 2$, we consider source data where an AVR was performed to be similar to the target data, and source data where an AVR was not performed to be dissimilar. We examine how augmenting the target task training data with different subsets of source data affects performance.

We define these subsets as follows. We denote by the superscript the criteria used to select each subset of the source data: +AVR denotes the patient received an AVR in combination with some other surgery, -AVR means the patient did not receive an AVR, and h or $\bar{h}$ denotes whether the patient received the surgery at the same hospital as the target task h or not. The subscript denotes the task of interest (i.e., which hospital h the target data are from). The subsets we consider in this chapter are

- $S_h^{+\text{AVR},h}$: the subset of source data available for task h from hospital h where patients received an AVR in addition to some other procedure(s).

- $S_h^{-\text{AVR},h}$: the subset of source data available for task h from hospital h where patients received a cardiac surgery, but did not receive an AVR.
- $S_h^h = S_h^{+\text{AVR},h} \cup S_h^{-\text{AVR},h}$: all available source data for task h from hospital h.

The relationship of these subsets at a single hospital is shown in Figure 3-1.

Hospital *h*

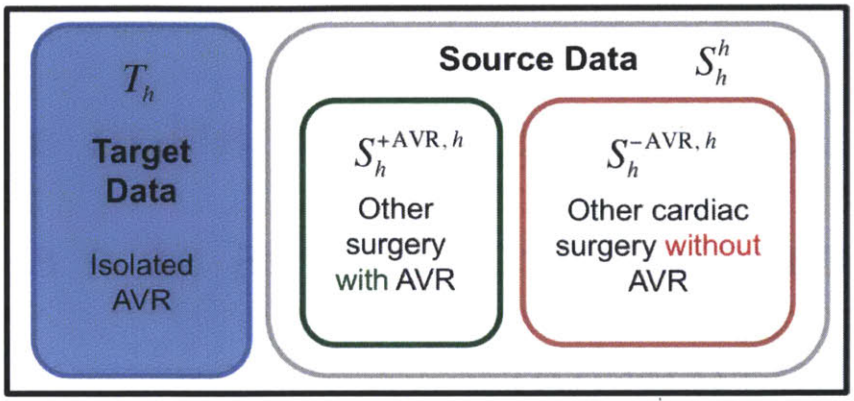

Figure 3-1: Diagram of the subsets of source data and target data at hospital h.

We hypothesize that because $S_h^{+\text{AVR},h}$ is most similar to the target data T_h, augmenting the target task training data with this set will improve performance on the target task more than other sets.

We compare against the baseline approach of using only target task training data to develop a model. We examine the tradeoff between amount of training data and how similar those data are to the target data.

3.1 Target-only

We first considered the baseline performance, where we trained a model using only the target training data and tested on the corresponding test set. We call this approach "target-only." The performance for both hospitals is shown in Table 3.1.

The AUC, precision, and recall for operative mortality for both hospitals were significantly better than random (p-value < 0.001). On the other hand, the performance on stroke for both hospitals was significantly worse than random (p-value $<$

Table 3.1: Target-only performance on Hospital 1 and Hospital 2 for Operative Mortality and Stroke. Training data used consists of 50% of the isolated AVRs from the hospital of interest. The average AUC, precision, and recall over the 100 test sets are reported, and the 95 % confidence intervals are included in parentheses.

		Hospital 1	**Hospital 2**
Operative Mortality	**AUC**	0.5787 (0.46, 0.71)	0.6601 (0.50, 0.75)
	Precision	0.0274 (0.01, 0.05)	0.0603 (0.03, 0.09)
	Recall	0.2800 (0.11, 0.56)	0.4291 (0.22, 0.61)
Stroke	**AUC**	0.4373 (0.29, 0.55)	0.4876 (0.39, 0.57)
	Precision	0.0183 (0, 0.04)	0.0245 (0.01, 0.04)
	Recall	0.1680 (0, 0.40)	0.1704 (0.04, 0.30)

0.05). This result is discouraging, but can be explained by the high-dimensionality of the problem and the small amount of relevant data.

The target-only model for Hospital 1 used only half of the available target data for training, i.e., just over 450 examples in each training set, with 9 or 10 adverse outcomes. This is an extremely small number of events considering that over 200 features were used to train each model. This ratio of training instances to features makes it extremely likely that the models will overfit to the training data. Thus, when the model is applied to a validation set, it is likely to demonstrate low performance. To verify that overfitting was the reason for the low performance on stroke, we compared the training AUC (where the model was tested on the sample used to develop it) with the test AUC over the 100 splits. The average difference between training and test AUC was 0.53 for the outcome of stroke, with 95% confidence interval (0.33, 0.72). This large discrepancy between the training and test performance demonstrates that overfitting (and thus, not enough training data) contributed to the poor performance.

Similarly, the performance for the outcome of stroke at Hospital 2 was below random. However, because there were more available target data from Hospital 2, the discrepancy between training and test AUCs was smaller than the discrepancy at Hospital 1 (mean difference in AUC of 0.46 and 95% confidence interval (0.33, 0.60)).

3.2 Using source data

We considered the following sets of source data: 1) $S_h^{+\text{AVR},h}$, 2) $S_h^{-\text{AVR},h}$, and 3) $S_h^h = S_h^{+\text{AVR},h} \cup S_h^{-\text{AVR},h}$.

We augmented our target task training data with each of these sets and evaluated performance on the test sets. We consider how variations in the training data affect performance of the models on the test data. For each outcome and each hospital, we show that more training data does not always result in better performance.

3.2.1 Outcomes in Hospital 1 and Hospital 2: Comparing Training and Test Sets

The size and number of adverse outcomes for each hospital in each subset of source data considered are shown in Table 3.2. The size of these subsets of source data vary, as well as the proportion of adverse outcomes that occur in them. The statistics for the target data are also shown for reference. The rates of adverse outcomes differ between procedure types as well as between hospitals.

Table 3.2: Distribution of outcomes in source data and target data for Hospital 1 and Hospital 2.

Hospital	**Data**	N	**Operative Mortality**	**Stroke**
1	$S_1^{+\text{AVR},1}$	1602	85 (5.3 %)	55 (3.4 %)
	$S_1^{-\text{AVR},1}$	7538	245 (3.3 %)	151 (2.0 %)
	S_1^1	9140	330 (3.6 %)	206 (2.3 %)
	T_1	917	19 (2.1 %)	21 (2.3 %)
2	$S_2^{+\text{AVR},2}$	2119	115 (5.4 %)	85 (4.0 %)
	$S_2^{-\text{AVR},2}$	8498	321 (3.8 %)	194 (2.3 %)
	S_2^2	10,617	436 (4.1 %)	279 (2.6 %)
	T_2	1602	45 (2.8 %)	46 (2.9 %)

3.2.2 Experimental Results

Tables 3.3 and 3.4 show the performance of models trained on augmented target task training data over the 100 test sets for the outcomes of operative mortality and stroke, respectively. These results demonstrate significant improvements in average performance over the target-only method, with p-values < 0.001.

For the outcome of operative mortality, augmenting the target training data with all of the available source data, S_h^h, resulted in significantly better AUC, precision, and recall than using either of the two smaller sets at both hospitals (p-values < 0.05). At Hospital 2, the subsets with more data outperformed the subsets with less ($S_h^{-\text{AVR},h}$ outperformed $S_h^{+\text{AVR},h}$, and S_h^h outperformed both of the other sets). These effects are what one might expect; more training data leads to better performance. However, the difference in performance between $S_h^{+\text{AVR},h}$ and $S_h^{-\text{AVR},h}$ was not significantly different at Hospital 1 for any of the three measures evaluated (p-values > 0.05). This is somewhat surprising, given that $S_1^{-\text{AVR},1}$ is almost 5 times as large as $S_1^{+\text{AVR},1}$ and contains approximately 3 times as many adverse events.

The observation that including more source data in training does not always lead to better performance is also true for the outcome of stroke. At Hospital 1, using $S_1^{+\text{AVR},1}$ resulted in significantly better performance than $S_1^{-\text{AVR},1}$ in AUC, precision, and recall (p-value < 0.05). Although using all of the source data still performed significantly better than using either of these subsets in terms of the AUC (p-value < 0.001), the precision and recall were not significantly different between using all of the data and using only $S_1^{+\text{AVR},1}$, the smallest of the three sets.

At Hospital 2, the results are similar; for the outcome of stroke, using $S_2^{-\text{AVR},2}$ resulted in significantly worse precision and recall and no significant change in the AUC compared to using $S_2^{+\text{AVR},2}$. For this task, using all of the available source data outperformed using either of the subsets in the AUC, but the precision and recall were significantly better when only $S_2^{+\text{AVR},2}$ was used.

In the target-only results (Table 3.1), the precision, recall, and AUC for the outcome of stroke were significantly worse than random. Augmenting the target training data with source data improved performance, and for the outcome of stroke, we were able to achieve performance significantly better than random in AUC, precision, and recall when we used $S_h^{+\text{AVR},h}$ or S_h^h (p-value < 0.05). Augmenting with the subset $S_h^{-\text{AVR},h}$, however, resulted in performance not significantly different from random at both hospitals. In this case as well, the set excluding surgeries where an AVR was performed had less utility.

In Tables 3.3 and 3.4, the maximum average AUC, precision, and recall for each hospital are in bold. Any values which were not significantly different from these maximum values are also in bold.

Table 3.3: Operative mortality: Average AUC, precision, and recall with 95% confidence intervals. Target task training data were augmented with different sets of source data from the target hospital.

Hospital	**Source**	**AUC**	**Precision**	**Recall**
	Data	(95 % CI)	(95 % CI)	(95 % CI)
1	$S_1^{+\text{AVR},1}$	0.6634 (0.52, 0.78)	0.0443 (0.02, 0.07)	0.4533 (0.22, 0.67)
	$S_1^{-\text{AVR},1}$	0.6670 (0.52, 0.78)	0.0447 (0.02, 0.07)	0.4567 (0.22, 0.67)
	S_1^1	**0.7038 (0.57, 0.83)**	**0.0470 (0.02, 0.07)**	**0.4800 (0.22, 0.67)**
2	$S_2^{+\text{AVR},2}$	0.7079 (0.62, 0.79)	0.0668 (0.04, 0.09)	0.4750 (0.30, 0.64)
	$S_2^{-\text{AVR},2}$	0.7274 (0.62, 0.81)	0.0741 (0.05, 0.09)	0.5271 (0.36, 0.68)
	S_2^2	**0.7409 (0.65, 0.81)**	**0.0793 (0.05, 0.10)**	**0.5636 (0.36, 0.70)**

Table 3.4: Stroke: Average AUC, precision, and recall with 95% confidence intervals. Target task training data were augmented with different sets of source data from the target hospital.

Hospital	**Source**	**AUC**	**Precision**	**Recall**
	Data	(95 % CI)	(95 % CI)	(95 % CI)
1	$S_1^{+\text{AVR},1}$	0.5932 (0.48, 0.69)	**0.0292 (0.01, 0.04)**	**0.2690 (0.10, 0.40)**
	$S_1^{-\text{AVR},1}$	0.5632 (0.47, 0.67)	0.0266 (0.01, 0.04)	0.2450 (0.10, 0.40)
	S_1^1	**0.6046 (0.49, 0.72)**	**0.0280 (0.01, 0.05)**	**0.2580 (0.10, 0.50)**
2	$S_2^{+\text{AVR},2}$	0.5407 (0.46, 0.62)	**0.0383 (0.02, 0.06)**	**0.2665 (0.13, 0.43)**
	$S_2^{-\text{AVR},2}$	0.5375 (0.45, 0.61)	0.0312 (0.01, 0.05)	0.2170 (0.09, 0.35)
	S_2^2	**0.5521 (0.46, 0.63)**	0.0343 (0.01, 0.06)	0.2387 (0.09, 0.39)

In this section, we showed that although $S_h^{-\text{AVR}}$ and S_h^h were larger than $S_h^{+\text{AVR}}$, they did not always result in better performance. In the following section, we will isolate the effects of similarity and size.

3.3 Isolating the effects of similarity and size

We isolated the effects of similarity and size by subsampling each training set to be the same size as the smallest ($S_h^{+\text{AVR},h}$). We used a two-tailed paired t-test and Wilcoxon signed-rank test to determine whether the performance using $S_h^{+\text{AVR},h}$ was significantly different from using the subsampled $S_h^{-\text{AVR},h}$ and S_h^h.

These results are shown in Tables 3.5 and 3.6. For both outcomes in hospital 1 and for stroke in hospital 2, the subsampled sets performed significantly worse than $S_h^{+\text{AVR},h}$. For operative mortality in hospital 2, however, the subsampled $S_2^{-\text{AVR},2}$ and the subsampled S_2^2 performed significantly better than $S_2^{+\text{AVR},2}$.

From the results shown in Tables 3.5 and 3.6, it appears that the extent to which the joint distributions of $S_h^{+\text{AVR},h}$ and $S_h^{-\text{AVR},h}$ are dissimilar differs between hospitals. Additionally, this dissimilarity affects performance differently at the two hospitals and for the different outcomes.

Table 3.5: Operative Mortality: Average AUC, precision, and recall with 95% confidence intervals. Training sets were randomly subsampled to be the same size as $S_h^{+\text{AVR},h}$.

Hospital	**Source Data**	**AUC**			**Precision**			**Recall**		
		95 % CI	t-test	signed-rank	95 % CI	t-test	signed-rank	95 % CI	t-test	signed-rank
1	$S_1^{+\text{AVR},1}$	**0.6634** **(0.52, 0.78)**			**0.0443** **(0.02, 0.07)**			**0.4533** **(0.22, 0.67)**		
	$S_1^{-\text{AVR},1}$	0.6269 (0.48, 0.76)	**<0.001**	**<0.001**	0.0378 (0.02, 0.07)	**<0.001**	**<0.001**	0.3867 (0.22, 0.67)	**<0.001**	**<0.001**
	S_1^1	0.6493 (0.50, 0.77)	**0.0291**	**0.0364**	0.0393 (0.02, 0.07)	**<0.001**	**<0.001**	0.4022 ((0.22, 0.67)	**<0.001**	**<0.001**
2	$S_2^{+\text{AVR},2}$	0.7079 (0.62, 0.79)			0.0668 (0.04, 0.09)			0.4750 (0.30, 0.64)		
	$S_2^{-\text{AVR},2}$	**0.7174** **(0.62, 0.79)**	**0.0053**	**0.0104**	**0.0731** **(0.04, 0.10)**	**<0.001**	**<0.001**	**0.5195** **(0.32, 0.70)**	**<0.001**	**<0.001**
	S_2^2	**0.7230** **(0.61, 0.79)**	**<0.001**	**<0.001**	**0.0715** **(0.04, 0.09)**	**<0.001**	**<0.001**	**0.5085** **(0.32, 0.68)**	**<0.001**	**<0.001**

Table 3.6: Stroke: Average AUC, precision, and recall with 95% confidence intervals. Training sets are randomly subsampled to be the same size as $S_h^{+\text{AVR},h}$.

Hospital	Source Data	AUC			Precision			Recall		
		95 % CI	*t*-test	signed-rank	95 % CI	*t*-test	signed-rank	95 % CI	*t*-test	signed-rank
1	$S_1^{+\text{AVR},1}$	**0.5932** **(0.48, 0.69)**	—	—	**0.0292** **(0.01, 0.04)**	—		**0.2690** **(0.10, 0.40)**	—	
	$S_1^{-\text{AVR},1}$	0.5104 (0.38, 0.62)	**<0.001**	**<0.001**	0.0227 (0, 0.04)	**<0.001**	**<0.001**	0.2090 (0, 0.40)	**<0.001**	**<0.001**
	S_1^1	0.5490 (0.41, 0.68)	**<0.001**	**<0.001**	0.0258 (0, 0.05)	**0.0295**	**0.0285**	0.2370 (0, 0.50)	**0.0295**	**0.0285**
2	$S_2^{+\text{AVR},2}$	**0.5407** **(0.46, 0.62)**	—		**0.0383** **(0.02, 0.06)**			**0.2665** **(0.13, 0.43)**	—	—
	$S_2^{-\text{AVR},2}$	0.5183 (0.40, 0.63)	**0.0010**	**0.0014**	0.0305 (0.01, 0.05)	**<0.001**	**<0.001**	0.2122 (0.09, 0.35)	**<0.001**	**<0.001**
	S_2^2	0.5143 (0.37, 0.62)	**<0.001**	**<0.001**	0.0304 (0.01, 0.06)	**<0.001**	**<0.001**	0.2117 (0.09, 0.39)	**<0.001**	**<0.001**

3.4 Summary

We demonstrated that subsets of the source data, selected based on whether or not the patient received an AVR, can significantly improve performance over the target-only approach. For operative mortality at Hospital 1, we achieved an AUC of 0.70, over a baseline of 0.58. Similarly, at hospital 2, the AUC improved from 0.66 to 0.74. On the outcome of stroke, we were able to improve performance from being significantly worse than random to significantly better than random (though still not good) at both hospitals. However, the subsets of source data were not equally helpful.

In this chapter, we evaluated similarity between source and target data based on whether or not an AVR was performed. We demonstrated that the extent to which examples in the source data are helpful depends on this criteria for similarity. It is thus important to consider the relationship of the source data to the target data. By subsampling, we isolated the effects of size and similarity and found that even when the sizes of the training sets were the same, performance differed significantly depending on the procedures allowed in the training set.

Chapter 4

Leveraging information across hospitals

In Chapter 3, we examined how including source information from other cardiac procedures at the same hospital could be useful when learning a model for our target task. We used the criteria of whether or not an AVR was performed to evaluate how similar source examples were to the target data. We found that incorporating other cardiac procedures, particularly those more related to the target data, could lead to significant improvements in performance. However, the data that were determined to be less related to the target data varied between hospitals and outcomes. For example, although adding $S_h^{-\text{AVR},h}$ to the source data detrimentally affected performance on stroke at Hospital 1, the same subset at Hospital 2 improved performance on operative mortality.

In this chapter, we investigate how leveraging data *across* hospitals can be used to learn a better model for our target task. To isolate the effect of transfer between institutions from the effect of transfer between procedures, we begin by considering only the isolated AVRs as available source data. For a target task h, this source data comes from the other hospital.

Next we again leverage subsets of source data that vary in whether or not patients received an AVR, as in Chapter 3, but across both institutions. Comparing the effects of these different methods of transfer allows us to examine the relative utility, in our data set, of transferring across institutions and transferring between different surgeries within an institution.

4.1 Setup

In each of the following sections in this chapter, we consider a subset of the available source data for each task h. The relationship of the source data S_h (blue) to the target data T_h (orange) is diagrammed in Figure 4-1.

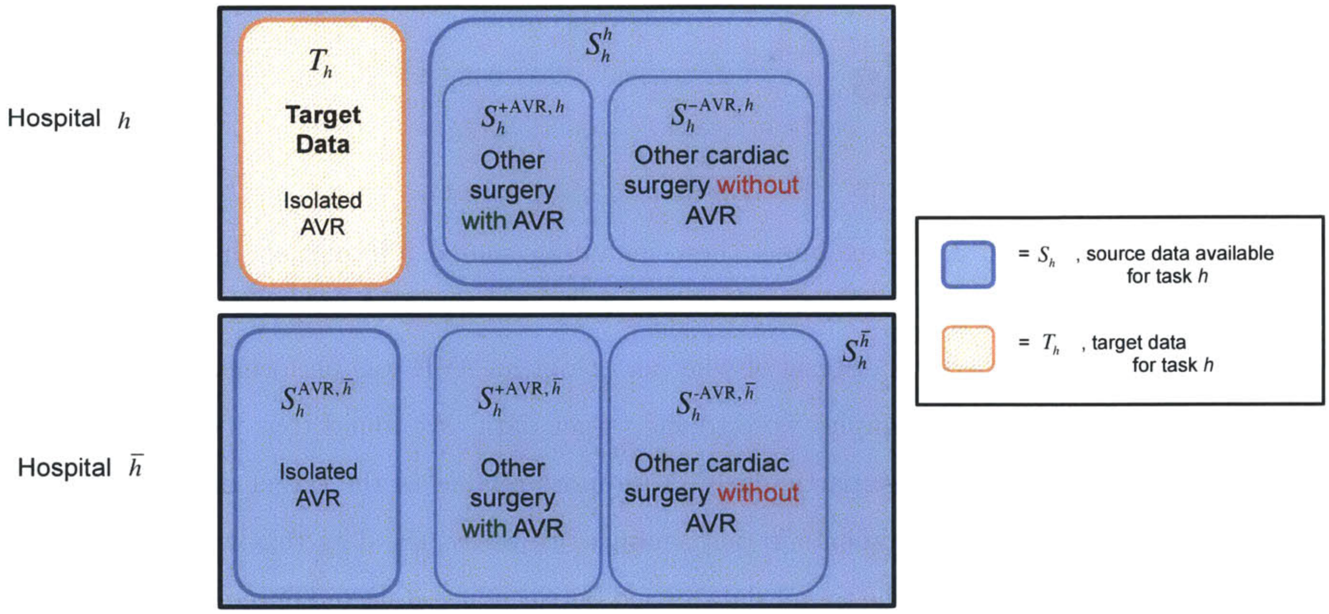
Hospital h
T_h
Target Data
Isolated AVR
S_h^h
$S_h^{+\mathrm{AVR},h}$
Other surgery with AVR
$S_h^{-\mathrm{AVR},h}$
Other cardiac surgery without AVR
Hospital $\bar{h}$
$S_h^{\mathrm{AVR},\bar{h}}$
Isolated AVR
$S_h^{+\mathrm{AVR},\bar{h}}$
Other surgery with AVR
$S_h^{-\mathrm{AVR},\bar{h}}$
Other cardiac surgery without AVR
$S_h^{\bar{h}}$
= S_h , source data available for task h
= T_h , target data for task h

Figure 4-1: Relationship of the available source data S_h with the target data, T_h.

4.2 Isolated Aortic Valve Replacements

We began by augmenting our target task training data, from hospital h, with the isolated AVR data from the other hospital ($\bar{h}$), which we denote as $S_h^{\mathrm{AVR},\bar{h}}$. We illustrate this in Figure 4-2. The blue shaded areas indicate the source data available which are not leveraged towards the task, while the unshaded area indicates source data that are used. We compare against the target-only results shown in Table 3.1.

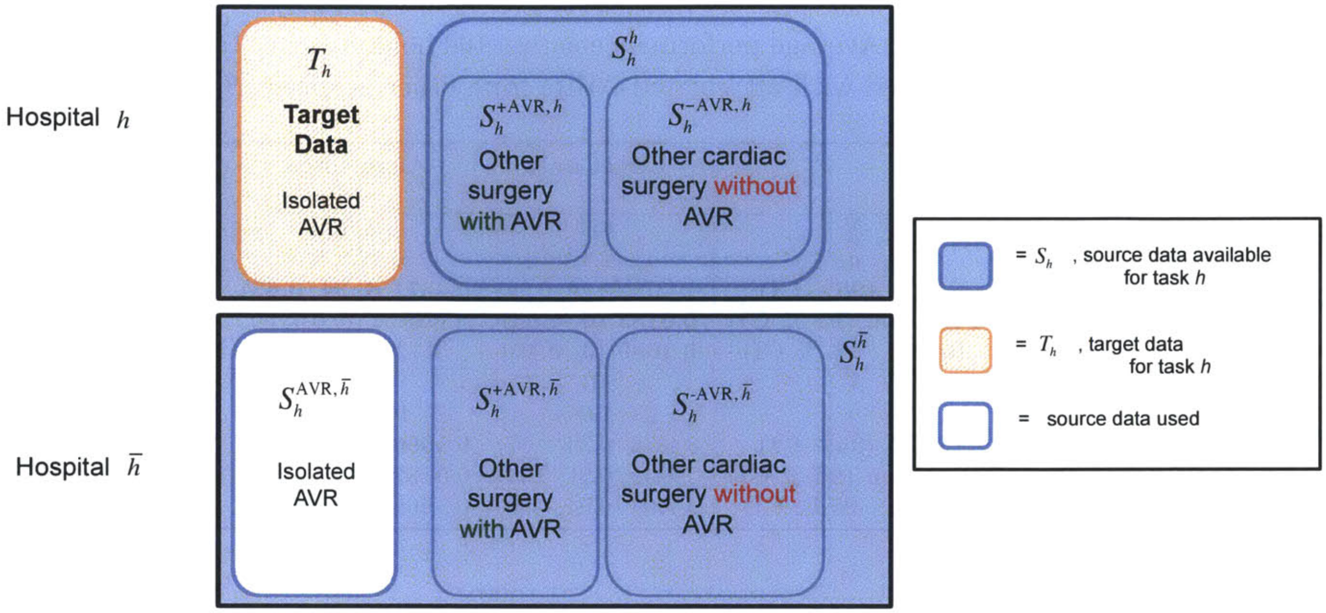

Figure 4-2: Leveraging data across hospitals for the same procedure.

Table 4.1: Operative Mortality: Average performance across 100 holdout sets. Target task training data are augmented with all isolated AVRs from other hospital ($S_h^{\mathrm{AVR},\bar{h}}$).

Hospital		Target-only	Augmented	p-value	
				t-test	sign-rank
	N	459	2061	—	—
	n	10 (2.2 %)	55 (2.7 %)	—	—
1	**AUC (95% CI)**	0.5787 (0.46, 0.71)	**0.6465 (0.51, 0.77)**	**<0.001**	**<0.001**
	Precision (95 % CI)	0.0274 (0.01, 0.05)	**0.0395 (0.02, 0.07)**	**<0.001**	**<0.001**
	Recall (95 % CI)	0.2800 (0.11, 0.56)	**0.4033 (0.22, 0.67)**	**<0.001**	**<0.001**
	N	801	1718	—	—
	n	22.5 (2.8 %)	41.5 (2.4 %)	—	—
2	**AUC (95% CI)**	0.6601 (0.50, 0.75)	**0.6915 (0.57, 0.78)**	**<0.001**	**<0.001**
	Precision (95 % CI)	0.0603 (0.03, 0.09)	**0.0671 (0.04, 0.09)**	**<0.001**	**<0.001**
	Recall (95 % CI)	0.4291 (0.22, 0.61)	**0.4771 (0.30, 0.64)**	**<0.001**	**<0.001**

For the outcome of operative mortality, all results for both hospitals were significantly better than target-only when the training data were augmented with the isolated AVRs from the other hospital (p-value < 0.001). This was similarly true for the outcome of stroke at Hospital 2. However, even though adding isolated AVRs from Hospital 2 to training when the target task was Hospital 1 also improved the AUC for the outcome of stroke, the performance was still worse than random (AUC

Table 4.2: Stroke: Average performance across 100 holdout sets. Target task training data are augmented with all isolated AVRs from other hospital ($S_h^{\mathrm{AVR},\bar{h}}$).

Hospital		**Target-only**	**Augmented**	**p-value**	
				t-test	sign-rank
1	N	459	2061	—	—
	n	11 (2.4 %)	57 (2.8 %)	—	—
	AUC (95% CI)	0.4373 (0.29, 0.55)	**0.4571 (0.34, 0.63)**	**0.0131**	**0.0258**
	Precision (95 % CI)	**0.0183 (0, 0.04)**	0.0137 (0, 0.03)	**0.0021**	**0.0026**
	Recall (95 % CI)	**0.1680 (0, 0.40)**	0.1260 (0, 0.30)	**0.0021**	**0.0026**
2	N	801	1718	—	—
	n	23 (2.9 %)	44 (2.6 %)	—	—
	AUC (95% CI)	0.4876 (0.39, 0.57)	**0.5066 (0.42, 0.59)**	**<0.001**	**<0.001**
	Precision (95 % CI)	0.0245 (0.01, 0.04)	**0.0282 (0.01, 0.05)**	**0.0048**	**0.0078**
	Recall (95 % CI)	0.1704 (0.04, 0.30)	**0.1961 (0.04, 0.35)**	**0.0048**	**0.0078**

< 0.5), and the precision and recall were significantly worse.

That adding this source data reduced both precision and recall initially surprised us. A possible explanation is that when we add the isolated AVRs from Hospital 2 to training for the target task of Hospital 1, 801 surgeries from Hospital 2 are added to the 459 available for training from Hospital 1. This means Hospital 2 is overrepresented compared to Hospital 1 in the training set. The model will try to fit the Hospital 2 data more than the Hospital 1 data because of their relative occurrence. If the features which contribute to stroke after isolated AVR are different for the two institutions, this will result in poor performance on Hospital 1.

To test this hypothesis, we subsampled the training data from Hospital 2 so that the number of examples and the number of adverse outcomes from Hospital 2 equalled the number from Hospital 1. The average AUC, precision, and recall after subsampling were no longer significantly different from the target-only values. By giving Hospital 1 and Hospital 2 fair representation in training, the model no longer overtrained to the examples from Hospital 2. However, even after subsampling, transferring information from Hospital 2 did not improve performance on Hospital 1.

Figure 4-3 compares the target-only performance against our experimental results from Chapter 3 (leveraging information from other procedures at the same hospital) and the results in Tables 4.1 and 4.2 (leveraging information from the same procedure at another hospital). Leveraging data from the same hospital but from different cardiac procedures (S_h^h) as described in Chapter 3 results in greater gains in performance.

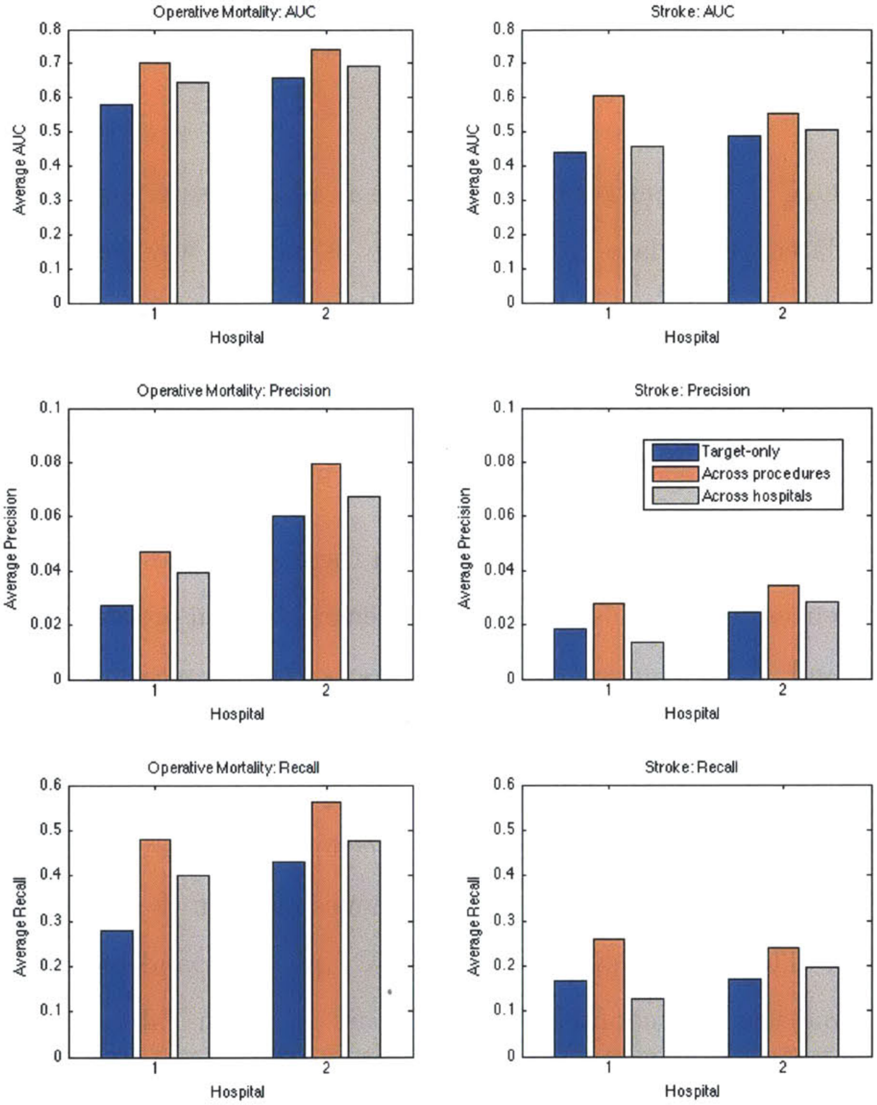

Figure 4-3: Comparing average performance using target-only, transfer across procedures, and transfer across hospitals on Hospitals 1 and 2 for outcomes of operative mortality (left) and stroke (right). For both outcomes at both hospitals, and for all performance metrics considered, leveraging other procedures at the same hospital outperforms the other approaches.

4.3 Other cardiac procedures

In addition to leveraging the information across hospitals while keeping the procedure constant, we also looked at how using examples for related procedures from *both* hospitals could improve performance. For hospital h, we considered the following subsets of source data, where $S_h^{+\text{AVR}} = S_h^{+\text{AVR},h} \cup S_h^{+\text{AVR},\bar{h}}$, and $S_h^{-\text{AVR}} = S_h^{-\text{AVR},h} \cup S_h^{-\text{AVR},\bar{h}}$:

- $S_h^{\text{AVR},\bar{h}} \cup S_h^{+\text{AVR}}$: all surgeries in S_h where an AVR was performed. This includes the isolated AVR data from hospital $\bar{h}$ as well as any instances from both hospitals where the patient received AVR and another surgery.
- $S_h^{-\text{AVR}}$: instances from both hospitals where the patient did not receive an AVR.
- $S_h = S_h^{\text{AVR},\bar{h}} \cup S_h^{+\text{AVR}} \cup S_h^{-\text{AVR}}$: the full set of available source data.

Figure 4-2 diagrams the relationships between these sets. These sets combine the effects of differences between hospitals and differences between procedures. The number of examples and adverse outcomes in each subset of the source data are shown in Table 3.2. We compare the results of the experiments using subsets of source data from both hospitals to the corresponding results in Chapter 3, where we used source data subsets that contained the same types of procedures, but only from the hospital of interest. We evaluate the difference in performance between these two approaches for each test set. The average differences in AUC, precision, and recall between these approaches and the 95% confidence intervals are shown in Tables 4.3 and 4.4. The full results for the experiments utilizing source data from both hospitals are in the Appendix (A.1-A.2).

In Chapter 3, we saw that when transferring information across procedures in a single hospital, different types of source data were not equally helpful (Tables 3.3-3.4). Similarly, the data from another hospital are not equally helpful and sometimes are not helpful at all. In the case of stroke, including Hospital 2 source data resulted in worse AUC, precision, and recall on Hospital 1 for all subsets of source data (Table 4.4). For operative mortality, including source data from another hospital

Table 4.3: Operative Mortality: Performance for each source data subset is compared to the corresponding results in Chapter 3, where the same type of source data were used, but only *from the same hospital.* A p-value < 0.05 indicates that utilizing the source data from another hospital (of the procedure subset considered) results in significantly different performance.

Hospital	**Source Data Comparison**	**AUC**			**Precision**			**Recall**		
		95 % CI	t-test	sign-rank	95 % CI	t-test	sign-rank	95 % CI	t-test	sign-rank
1	$(S_1^{\text{AVR},2} \cup S_1^{+\text{AVR}}) - S_1^{+\text{AVR},1}$	-0.0059 (-0.10, 0.09)	0.2596	0.1576	**-0.0126** **(-0.03, 0.01)**	**<0.001**	**<0.001**	**-0.1289** **(-0.33, 0.11)**	**<0.001**	**<0.001**
	$S_1^{-\text{AVR}} - S_1^{-\text{AVR},1}$	**0.0174** **(-0.06, 0.07)**	**<0.001**	**<0.001**	0.0003 (-0.02, 0.01)	0.7437	0.8265	0.0033 (-0.22, 0.11)	0.7437	0.8265
	$S_1 - S_1^1$	**0.0141** **(-0.04, 0.08)**	**<0.001**	**<0.001**	-0.0008 (-0.01, 0.01)	0.3197	0.3173	-0.0078 (-0.11, 0.11)	0.3197	0.3173
2	$(S_2^{\text{AVR},1} \cup S_2^{+\text{AVR}}) - S_2^{+\text{AVR},2}$	**0.0292** **(0, 0.06)**	**<0.001**	**<0.001**	**0.0102** **(-0.01, 0.03)**	**<0.001**	**<0.001**	**0.0724** **(-0.04, 0.18)**	**<0.001**	**<0.001**
	$S_2^{-\text{AVR}} - S_2^{-\text{AVR},2}$	0.0030 (-0.04, 0.04)	0.1373	0.2058	0.0010 (-0.01, 0.02)	0.1450	0.1892	0.0071 (-0.09, 0.13)	0.1493	0.1864
	$S_2 - S_2^2$	-0.0004 (-0.03, 0.04)	0.8343	0.6951	**-0.0042** **(-0.03, 0.01)**	**<0.001**	**<0.001**	**-0.0303** **(-0.17, 0.09)**	**<0.001**	**<0.001**

Table 4.4: Stroke: Performance for each source data subset is compared to the corresponding results in Chapter 3, where the same type of source data were used, but only *from the same hospital.* A p-value < 0.05 indicates that utilizing the source data from another hospital (of the procedure subset considered) results in significantly different performance.

Hospital	**Source Data Comparison**	**AUC**			**Precision**			**Recall**		
		95 % CI	t-test	sign-rank	95 % CI	t-test	sign-rank	95 % CI	t-test	sign-rank
1	$(S_1^{\text{AVR},2} \cup S_1^{+\text{AVR}}) - S_1^{+\text{AVR},1}$	**-0.0141** **(-0.09, 0.07)**	**0.0010**	**0.0017**	**-0.0104** **(-0.03, 0.01)**	**<0.001**	**<0.001**	**-0.0960** **(-0.30, 0.10)**	**<0.001**	**<0.001**
	$S_1^{-\text{AVR}} - S_1^{-\text{AVR},1}$	**-0.0500** **(-0.12, 0.02)**	**<0.001**	**<0.001**	**-0.0118** **(-0.03, 0.01)**	**<0.001**	**<0.001**	**-0.1090** **(-0.30, 0.10)**	**<0.001**	**<0.001**
	$S_1 - S_1^1$	**-0.0578** **(-0.13, 0.02)**	**<0.001**	**<0.001**	**-0.0125** **(-0.03, 0.01)**	**<0.001**	**<0.001**	**-0.1150** **(-0.30, 0.10)**	**<0.001**	**<0.001**
2	$(S_2^{\text{AVR},1} \cup S_2^{+\text{AVR}}) - S_2^{+\text{AVR},2}$	0.0070 (-0.09, 0.15)	0.2074	0.3658	**-0.0111** **(-0.04, 0.01)**	**<0.001**	**<0.001**	**-0.0774** **(-0.26, 0.09)**	**<0.001**	**<0.001**
	$S_2^{-\text{AVR}} - S_2^{-\text{AVR},2}$	**0.0145** **(-0.02, 0.05)**	**<0.001**	**<0.001**	**0.0057** **(-0.01, 0.02)**	**<0.001**	**<0.001**	**0.04** **(-0.09, 0.13)**	**<0.001**	**<0.001**
	$S_2 - S_2^2$	**0.0091** **(-0.03, 0.04)**	**<0.001**	**<0.001**	**-0.0048** **(-0.03, 0.01)**	**<0.001**	**<0.001**	**-0.0330** **(-0.17, 0.09)**	**<0.001**	**<0.001**

negatively affected the performance when the subset contained only +AVR (patients who received AVR in conjunction with some other procedure(s)).

However, including data from another hospital can be helpful. For operative mortality, adding source data from Hospital 2 either improved the AUC significantly on Hospital 1 or had no significant effect. For the cases where the AUC was significantly improved, the precision and recall were not significantly different. Although the performance does not significantly increase for all measures considered, these results show that the data from another hospital can be useful.

In all cases, adding Hospital 2 to the available source data resulted in larger magnitude differences on Hospital 1 than adding Hospital 1 data for Hospital 2. In section 4.2, we showed that when Hospital 2 was overrepresented in the training set compared to Hospital 1, it resulted in negative transfer. When all of the source data from Hospital 2 are used, Hospital 2 contributes 1.8 times the number of examples Hospital 1 contributes to the training data, and twice as many adverse outcomes. The effect of source data has been shown in previous work to be larger in cases where less target data is available in other applications [12, 19]. Additionally, this has been shown to be specifically true for transfer between institutions for clinical risk-stratification tools [3].

4.4 Summary

In this chapter, we explored transfer of knowledge across institutions. We compared the results from this type of transfer with our results from intra-institution transfer in Chapter 3.

We reaffirmed that the effects of adding source data vary across institutions and across outcomes. We have demonstrated that using all of the data is not always the best option. In some cases, it can result in significantly worse performance. Including source information in training that results in a negative effect on performance on the target task is called "negative transfer" [1]. It is important to account for differences in similarity between source and target examples when determining what subset of the available source data to use in training. We have thus far only explored a heuristic for deciding what the best methods of transfer might be (e.g., procedures "similar" to isolated AVRs, procedures from the same hospital). Previous work has explored automated approaches to exploiting the positive effects of utilizing source data while avoiding the negative ones [12, 19]. We will investigate these methods further in the next chapter.

Chapter 5

Automatic identification of relevant training data: leveraging data across procedures and hospitals

In the previous chapters, we demonstrated that source data from other hospitals and procedures can be leveraged in training to improve performance on our target task. The extent to which the source data help depends on how related they are to the target task, and in some cases, adding source data resulted in *worse* performance. This suggests that better performance could be achieved by removing examples which cause negative transfer.

To address this problem, we seek to develop an automated method for identifying relevant training data from the full set of available source data for each task h, S_h. We first propose a metric for evaluating similarity of source data to target data. This metric can vary according to 1) what space it is measured in, and 2) how it is measured (e.g., squared Euclidean distance). We will evaluate the squared Euclidean distance in all cases and focus on looking at different subspaces $\mathcal{X}'$ for computing the distance. We describe a selection algorithm that removes dissimilar data from the training set. Finally, we propose an instance-weighting scheme, where no data are removed from training but each training example is assigned a different weight. While removing data requires selecting a reasonable cutoff, this final algorithm does not rely on any user-selected hyperparameters and is therefore fully automated.

5.1 Feature space: $\mathcal{X}'$

In Chapters 3 and 4, we chose a reduced feature space in which to evaluate similarity of source and target data. This feature space contained two features: 1) which hospital the surgery was performed at, and 2) whether or not an AVR was performed. However, these features are not sufficient to capture all of the relevant differences between source and target examples. To address this problem, we examine two feature spaces which encompass more features. We hypothesize that the relevant differences between the source and target data can be better captured in these spaces.

The feature spaces we consider are as follows:

- **Full feature space, $\mathcal{X}$**: We first evaluated distance of source data from target data using the full feature space. To maintain consistency with our previous methods, we evaluated this distance before discretizing continuous features. In our past experiments, quintiles for discretization were determined based on the training data. Because the distance metric is used to determine which data to include in training, we first computed the distance to determine which data to remove. We then discretized based on the included training examples.

  Missing data for continuous features were replaced with the mean, and features were normalized to have a range between 0 and 1 by subtracting the minimum value of the feature and dividing by the range. Useful source data were selected according to how far each example was from the mean of the target training data, based on the squared Euclidean distance. This feature space consisted of 223 features.

- **Procedural variables, $\mathcal{X}^{\text{procedures}}$**: We used all 39 binary features that contained information about which *procedures* the patient received. This allowed more variation in distance between source and target data than only considering the single procedural variable of whether or not an AVR was performed. In this space, all target data had the same feature vector as the target mean.

We will describe our algorithms and results using the full feature space, $\mathcal{X}$, in depth. We compare these results to when the reduced space $\mathcal{X}^{\text{procedures}}$ is used.

5.2 Selection algorithm

For each of the 100 training-test splits, we computed the mean target data instance using the target task training data. We then computed the squared Euclidean distance of each example in the training set (target and source) to this mean. The distance of each example from this target mean gives us a sense of how dissimilar it is to the target data. Because the source data are not drawn from the same distribution as the target data, we expect the distribution of source data distances to be shifted right relative to the distribution of target data distances.

An example of this is shown in Figure 5-1. To select relevant source data, we considered two criteria: 1) percentile cutoffs of *target data* distance from the target mean (i.e., according to the histogram in the top panel in Figure 5-1), and 2) percentile cutoffs of *source data* distance from the target mean (i.e., according to the histogram in the bottom panel in Figure 5-1). Source data further than a specified cutoff (e.g., 90th percentile of source data distances) were removed from the training set. No target task training examples were removed. Our hypothesis is that source examples further away from the target data mean bear less resemblance to the target data. We expect that at as we increase the cutoff (allow source data further and further from the target mean), the utility of the data included will decrease. Including these data may have no effect on performance, or it may hurt performance.

This selection algorithm was unsupervised; we did not utilize the outcomes of any of the patients to choose relevant data.

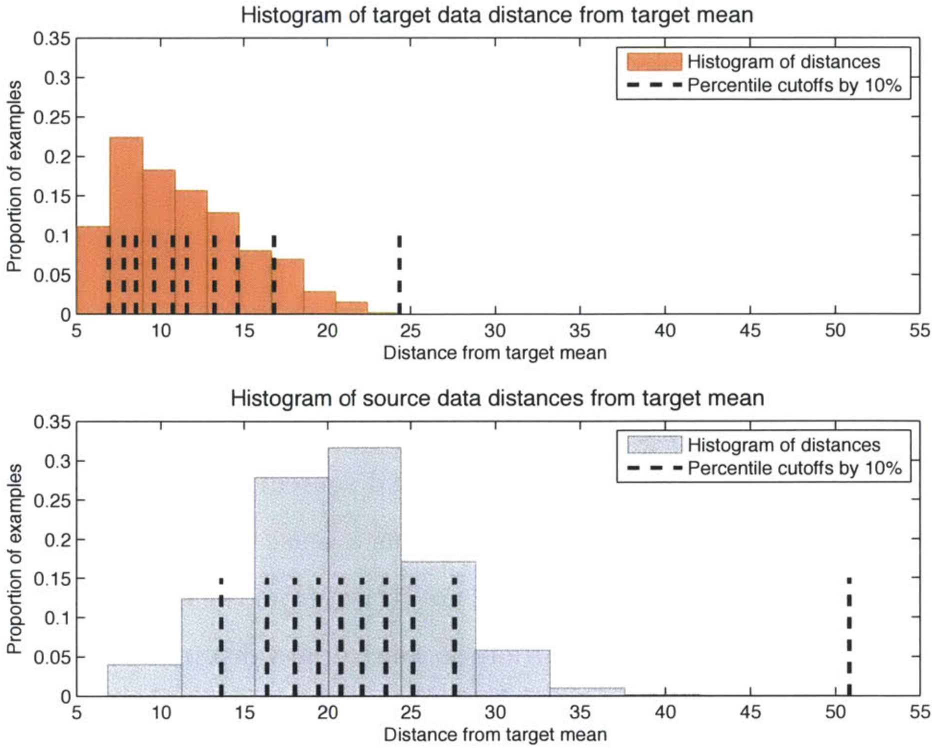

Figure 5-1: Histograms of distances of target (top) and source (bottom) data from target data mean. Percentile cutoffs of target and source data distance are shown in the respective histograms. Cutoffs based on source data distance allow for including source data outside the radius of the furthest target task example, whereas cutoffs based on the target data do not.

5.3 Instance-weighting

We also considered an instance-weighting scheme based on the distance metric we used in the selection algorithm. Rather than removing examples, we used the distance from the target mean to weight examples as more or less important in training. We used the instance-weighting option in the LIBLINEAR package [49]. The diagram in Figure 5-2 demonstrates how we constructed weights I_i for each example i, using the squared Euclidean distance d_i as in the previous section.

These weights were calculated for both source and target training instances. We constructed these weights so that any example at least as close to the target mean as the furthest target example had a weight greater than 1. Instances outside of

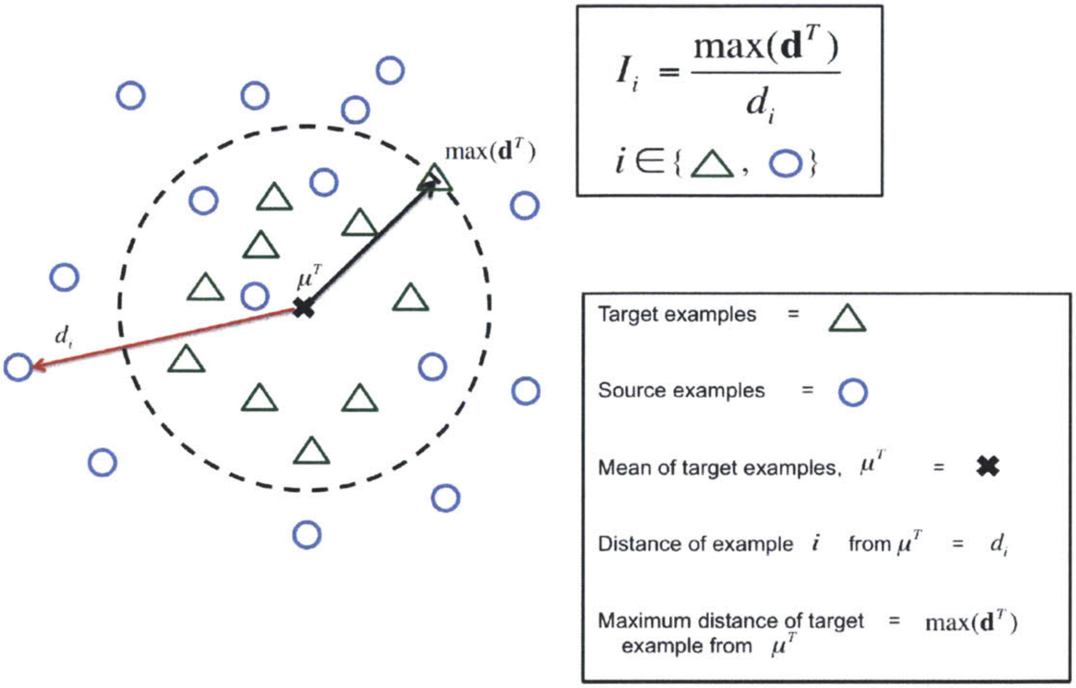
$I_i = \frac{\max(\mathbf{d}^T)}{d_i}$
$i \in \{\triangle, \bigcirc\}$
$\max(\mathbf{d}^T)$
μ^T
d_i
Target examples = △
Source examples = ○
Mean of target examples, μ^T = ✖
Distance of example i from μ^T = d_i
Maximum distance of target example from μ^T = $\max(\mathbf{d}^T)$

Figure 5-2: Diagram of distance metric and instance-weighting scheme. Distances (d_i) of all source and target examples from the target mean (μ^T) are computed. Source examples which are further from the target mean than the furthest target example ($\max(\mathbf{d}^T)$) will have a weight less than 1, and any examples inside this radius will have a weight greater than 1.

this radius had weights less than 1. This weighting scheme reflects our intuition that source examples that fall within the bounds of the target data (between the target mean and the the most extreme target example) are more similar to the target data than examples that fall outside of these bounds. Additionally, it addresses the scenario where there are outliers in the target data; the distances of these outliers from the target mean will be further than the distances of non-outliers, and they will therefore be weighted as less important.

5.4 Experimental Results

In this section, we show experimental results for each of the methods described above. We compare the instance-weighting scheme to choosing subsets of source data based on distance cutoffs. We explore two sets of cutoffs, each based on the distribution of distances from the target mean. First, we explore cutoffs based on the *target data* distribution of distances from the target mean. Second, we explore cutoffs based on the *source data* distribution of distances from the target mean. An example of the difference between these distributions is shown in Figure 5-1. The cutoffs range from the 10th to the 100th percentiles in increments of 10.

Figures 5-3 and 5-4 compare the average AUC when instance-weighting was used against the average AUC when selecting subsets of source data according to distance from the target mean (in the full feature space). The green line shows the average AUC when the percentile cutoffs are selected according to the *target data* distances (top panel in Figure 5-1), and the blue line shows the average AUC when the percentile cutoffs are selected based on the *source data* distances (bottom panel in Figure 5-1). Because the 100th percentile of the target task distances does not encompass the 100th percentile of the source data distances, the green line does not contain the point where all available source data are used.

As more training data are added, performance generally improves. In some cases, using all of the available data results in worse performance than using some subset. This is most obvious in the case of stroke at Hospital 1 (left panel of Figure 5-4), where using all of the data results in an average AUC of 0.55 but using only a subset results in a significantly higher average AUC of 0.57, as well as significantly higher average precision and recall (Table 5.2). Additionally, source examples that lie further from the target mean than the furthest target example still have utility.

The instance-weighting approach either demonstrated significant improvements over simply using all of the data without weights, or it demonstrated no significant differences. For Hospital 1, the average AUC for operative mortality when the data were weighted was significantly better than when they were not, and the precision and recall were not significantly different. For Hospital 2, all results for operative mortality were significantly improved over using all of the data without instance-weighting. In the case of stroke, instance-weighting provided no significant benefits over using all of the source data without weights, but it was not significantly worse, either.

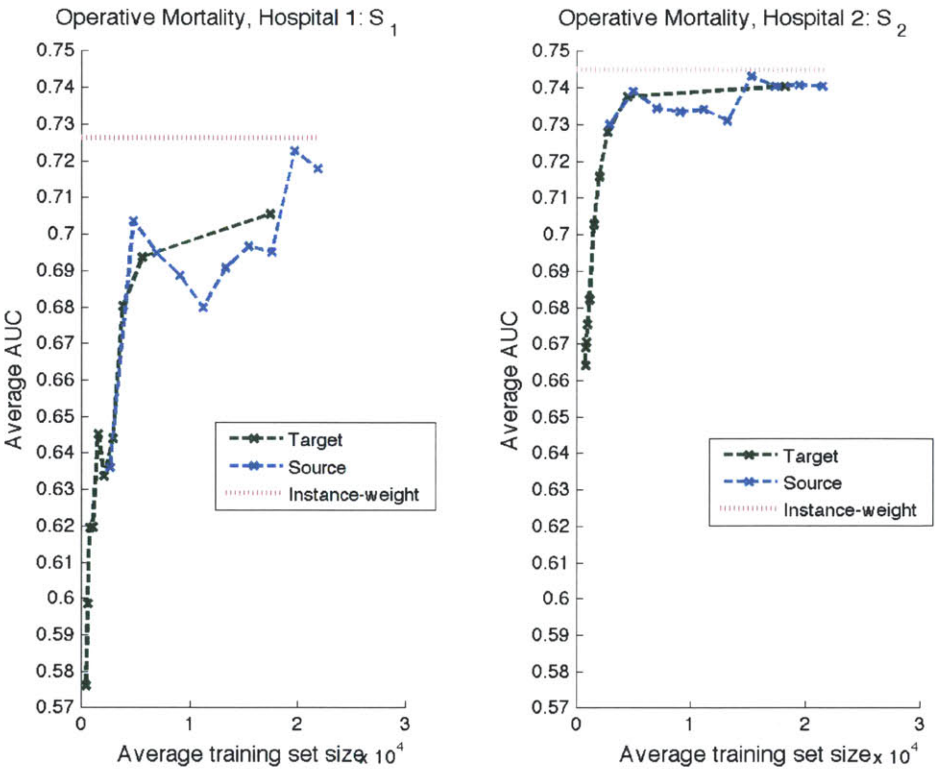

Figure 5-3: S_h in $\mathcal{X}$, Operative Mortality

As in the results shown in Chapter 4, when the available data from Hospital 2 were added to training for a task at Hospital 1, there was a large difference in performance compared to when Hospital 2 data were not used. In the case of operative mortality, the average AUC for the instance-weighting approach increased from 0.70 to 0.73, but the average precision and recall stayed approximately the same (Table 5.1). The performance for stroke dropped significantly, from 0.61 to 0.55 (Table 5.2), as in Chapter 4. Additionally, the recall for the outcome of stroke dropped from 0.29 to 0.14, which is well below the recall for a random classifier (0.20). Adding source data from Hospital 1 did not result in a large change in the magnitude of the performance metrics for either outcome at Hospital 2.

As we showed in Chapter 4, this difference in transfer effect between institutions is due to the difference in amount of target data available to each task. There are roughly twice as many target examples from Hospital 2 as there are from Hospital 1. A

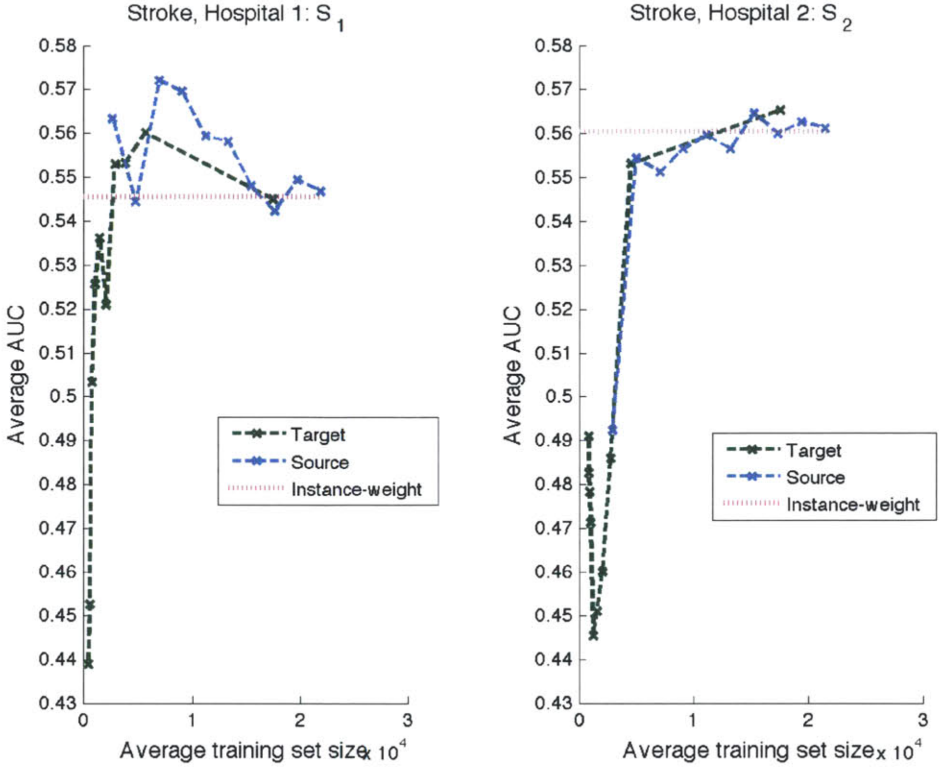

Figure 5-4: S_h in $\mathcal{X}$, Stroke

distance metric that better accounts for hospital-specific differences could help avoid the negative transfer between institutions. When negative transfer occurred, a subset of S_h was able to achieve significantly better performance than instance-weighting (e.g., stroke in Hospital 1). Because selecting a subset is equivalent to applying a weight of 0 to "remove" examples, modifying the instance-weighting function to something like an L1-norm could help improve performance.

Finally, in all of our experiments, we have observed that the performance for operative mortality is much higher than for stroke, despite the comparable number of events that occur in the training sets. We hypothesize that this difference in performance is due to the features we use to learn the model. As discussed in Chapter 2, we consider only preoperative features recorded in the STS database, and we do not consider any nonlinear relationships between features outside of discretizing the continuous features. It is possible that either risk of stroke is harder to explain using only preoperative features than risk of death, or that there are more complex, nonlinear relationships between the features and the outcome of stroke that we have not considered in this work.

Table 5.1: Operative Mortality: Instance-weighting and data selection when available source data are from the both hospitals (S_h) using distance metric in $\mathcal{X}$.

Hospital	**Performance**	**Weighted**	S_h^h	**p-value**		**Best subset**	**p-value**	
	Metric	S_h^h		t-test	signed-rank	of S_h^h	t-test	signed-rank
1	**AUC**	**0.7263**	0.7179	**<0.001**	**<0.001**	0.7228	0.1045	0.1586
		(0.60, 0.84)	(0.59, 0.83)			(0.62, 0.83)		
	Precision	0.0466	0.0462	0.1583	0.2891	0.0466	1	1
		(0.02, 0.07)	(0.02, 0.07)			(0.02, 0.07)		
	Recall	0.4767	0.4722	0.1583	0.2891	0.4767	1	1
		(0.22, 0.67)	(0.22, 0.67)			(0.22, 0.67)		
2	**AUC**	**0.7449**	0.7405	**<0.001**	**<0.001**	0.7432	0.0872	0.0617
		(0.66, 0.82)	(0.66, 0.81)			(0.66, 0.82)		
	Precision	**0.0771**	0.0750	**0.0019**	**0.0016**	0.0772	0.9356	0.8848
		(0.06, 0.10)	(0.05, 0.10)			(0.06, 0.10)		
	Recall	**0.5484**	0.5332	**0.0018**	**0.0036**	0.5488	0.9368	0.9175
		(0.39, 0.70)	(0.35, 0.70)			(0.39, 0.70)		

Table 5.2: Stroke: Instance-weighting and data selection when available source data are from both hospitals (S_h) using distance metric in $\mathcal{X}$.

Hospital	**Performance**	**Weighted**	S_h^h	**p-value**		**Best subset**	**p-value**	
	Metric	S_h^h		t-test	signed-rank	of S_h^h	t-test	signed-rank
1	**AUC**	0.5454	0.5468	0.3209	0.2281	0.5721	$<$ **0.001**	$<$ **0.001**
		(0.43, 0.64)	(0.44, 0.64)			(0.46, 0.67)		
	Precision	0.0149	0.0155	0.4997	0.5312	0.0218	$<$ **0.001**	$<$ **0.001**
		(0, 0.04)	(0, 0.03)			(0, 0.03)		
	Recall	0.1370	0.1430	0.4997	0.5312	0.2010	$<$ **0.001**	$<$ **0.001**
		(0, 0.40)	(0, 0.30)			(0, 0.30)		
2	**AUC**	0.5604	0.5612	0.5012	0.5103	0.5654	**0.0240**	0.0697
		(0.48, 0.64)	(0.48, 0.63)			(0.49, 0.64)		
	Precision	0.0281	0.0296	0.0511	**0.0405**	0.0277	0.7450	0.5492
		(0.01, 0.04)	(0.01, 0.05)			(0.01, 0.04)		
	Recall	0.1952	0.2057	0.0511	**0.0405**	0.1930	0.7450	0.5492
		(0.09, 0.30)	(0.09, 0.35)			(0.09, 0.30)		

5.5 Comparing Methods in Different Feature Spaces

In this section, we compare the results from using $\mathcal{X}^{\text{procedures}}$ with using the full feature space $\mathcal{X}$.

Table 5.3 shows the average difference in performance between the full space and the reduced procedural space, the 95% confidence intervals, and the two-tailed *t*-test and Wilcoxon signed-rank test *p*-values when instance-weighting is used and source data from both hospitals are available.

Table 5.3: Comparing feature spaces to compute distance in when available data are from S_h. Models were trained using instance-weighting based on the squared Euclidean distance in $\mathcal{X}^{\text{full}}$ and $\mathcal{X}^{\text{procedures}}$. The average difference in performance between the two feature spaces is shown along with the 95% confidence interval.).

		Operative Mortality			**Stroke**		
Hospital	**Performance Metric**	**Difference**	**p-value**		**Difference**	**p-value**	
			t-test	signed-rank		*t*-test	signed-rank
1	**AUC**	**0.0076** (-0.01, 0.03)	**<0.001**	**<0.001**	**0.0095** (-0.01, 0.04)	**<0.001**	**<0.001**
	Precision	0.0002 (-0.01, 0.01)	0.5955	0.7905	0.0004 (-0.01, 0.02)	0.6397	0.6346
	Recall	0.0022 (-0.11, 0.11)	0.5955	0.7905	0.0040 (-0.10, 0.20)	0.6397	0.6346
2	**AUC**	**-0.0012** **(-0.01, 0.01)**	**0.0031**	**0.0238**	0.0010 (-0.03, 0.03)	0.4527	0.5004
	Precision	-0.0009 (-0.01, 0.01)	0.1540	0.1480	**0.0019** **(-0.01, 0.01)**	**0.0040**	**0.0043**
	Recall	-0.0062 (-0.09, 0.09)	0.1589	0.4863	**0.0135** **(-0.09, 0.09)**	**0.0040**	**0.0043**

For the outcome of operative mortality, $\mathcal{X}^{\text{procedures}}$ did not demonstrate consistently significant differences for all performance metrics from the full feature space for either hospital. Although there were some significant differences in the AUC, none of the differences in precision or recall were significant. The performance when we use the procedural feature space to calculate the distance from the target mean is not very different from when we use the full feature space.

For the outcome of stroke at Hospital 1, the performance metrics also did not demonstrate consistent differences between the full feature space and the procedural space. However, at Hospital 2, the full space tends to outperform the procedural space (Table 5.3).

5.6 Summary

Instance-weighting is an automated approach that can result in a small, but significantly better performance than using all of the source data without weights. We have explored this approach in two feature spaces and used weights based on the squared Euclidean distance in these feature spaces.

Instance-weighting did not help counter the negative transfer that occurred when source data from Hospital 2 were used in training for the target task of stroke at Hospital 1 (Table 5.2). However, in the cases where negative transfer occurred, selecting a subset of S_h was able to achieve significantly better performance than instance-weighting. These results suggest that by modifying the instance-weighting function, we can better avoid negative transfer. Additionally, much smaller feature subspaces, such as the procedural one we considered, can achieve similar performance to the full feature space in some cases.

In summary, instance-weighting is a promising method to take better advantage of the useful instances in the source data. However, it was unable to avoid the negative transfer that occurred between institutions. Finding a more suitable feature space in which to evaluate this distance and considering other functions of the distance as instance-weights will give us more insight into when this method works best.

Chapter 6

Conclusions and Future Work

6.1 Summary

Clinical risk-stratification tools are often developed for specific patient populations. After applying a number of exclusion criteria, the data available for developing and testing these models may be insufficient. In cardiac surgery, models such as the STS risk models group types of surgeries using expert knowledge of related procedures, while models such as the EuroSCORE group all surgeries together. Both models are global in that they do not account for institution-specific differences, and the EuroSCORE models are additionally global in the sense that they do not account for procedure-specific differences.

In this thesis, we developed risk-stratification tools for patients at individual hospitals for outcomes of operative mortality and stroke after isolated aortic valve replacement. We began with a "target-only" approach, which used only the isolated AVR data from the hospital of interest. We demonstrated that using only target data to develop and test a model resulted in unsatisfactory AUC, precision, and recall for both outcomes at both hospitals.

To address the problem of too little data, we investigated instance-transfer methods in the field of transfer learning. Transfer learning assumes that the source data and the target data have related joint distributions of the features and outcome; if this is not the case, using the source data can negatively affect performance.

We hypothesized that leveraging available source data from both hospitals could improve performance on risk-stratifying isolated AVRs, but that instances more similar to the target data would be more helpful. To test this hypothesis, we addressed the question of how to evaluate similarity between source and target examples. In Chapters 3 and 4, we demonstrated that similarity of source and target data, evaluated based on the procedure(s) the patients received and the hospital the surgery was performed at, affects performance.

In Chapter 3, we augmented target task training data with available source data from the target hospital. We considered 3 subsets of the available source data that varied in how similar they were to the target data based on the criteria of whether or not an AVR was performed ($S_h^{+\text{AVR}}$, $S_h^{-\text{AVR}}$, and S_h^h). We showed that when these subsets were subsampled to the same size, augmenting target task training data with the subset containing examples where patients received an AVR with some other procedure (+AVR) resulted in the best AUC, precision, and recall for both outcomes at Hospital 1, and for stroke at Hospital 2. However, using all of the available source data still resulted in the best AUC for both outcomes at both hospitals. For Hospital 1, we achieved an average AUC of 0.70 for operative mortality, compared to the target-only performance of 0.58. For Hospital 2, leveraging all of the available source data resulted in an average AUC of 0.74, compared to the target-only performance of 0.66. For the outcome of stroke, utilizing source data from the same hospital in training also led to significantly improved performance over the target-only baseline for both hospitals.

In Chapter 4, we also considered the available source data from the other hospital. We showed that adding data from another hospital could result in negative transfer. These effects were more significant for Hospital 1, because it had less available target data. However, leveraging the data from the other hospital also had positive effects; at Hospital 1, using source data from Hospital 2 increased performance from an average AUC of 0.70 to 0.72 for the outcome of operative mortality.

Finally, in Chapter 5, we investigated automated approaches to leveraging auxiliary data. We showed that weighting instances of the available source data can achieve comparable or significantly better performance than using the available source data without weights. For the outcome of mortality at Hospital 1, instance-weighting the available source data (S_1) resulted in an average AUC of 0.73, over an average AUC of 0.72 when instance-weighting was not used. However, the improvements in performance we saw were small.

There are several ways we believe performance can be improved. Throughout this thesis, we have observed that the performance of our models for operative mortality has been significantly better than the performance of our models for stroke. We hypothesize that this could be because intraoperative features contribute more to predicting risk of stroke than preoperative features, or because there are nonlinear relationships between the features and stroke that we have not considered here.

Another area for further research is how to avoid negative transfer. In Chapters 4 and 5, we saw that utilizing data from the other hospital ($\bar{h}$) resulted in negative transfer at hospital h. Our automated approaches to selecting and weighting training data were not able to avoid this effect. This effect was stronger for Hospital 1 than for Hospital 2, because Hospital 1 had much less target data than Hospital 2.

6.2 Future Work

To address the problems noted above, we would like to investigate the following in our future work:

- **Dimensionality Reduction.** As we showed in Chapter 5, a reduced feature space such as $\mathcal{X}^{\text{procedures}}$ was able to achieve performance comparable to the full feature space $\mathcal{X}$ in some cases. We hypothesize that not all features we consider are important in distinguishing useful source data from harmful source data. Exploring ways to determine which feature subspace best characterizes the similarity between source and target examples is important to improving the instance weights we used in this work.

- **Considering other functions to compute instance weights.** The weights we use currently are developed according to an idea of how the target data and source data are related; source data that are contained within the "radius" of target examples are viewed as more similar and should be given a weight greater than 1. Examples outside of this radius should have a weight less than 1. However, the weights we use are not the only way to achieve this relationship. Computing instance-weights using other functions of the distance may lead to better results. For example, we could use a weighting function which also gives outliers in the target data a weight less than 1.

- **Identifying clusters of patients**. Although we have quantified distance of source data from target data using the mean of the target data, this could be generalized to looking at several clusters of target data. If there exists structure within the target data, clustering the examples and then determining the distance of the source data to these cluster centroids could be a more effective way of deciding which source data to incorporate.

- **Considering the utility of incorporating intraoperative features.** In this thesis, we showed that the performance of our models for operative mortality was much better than the performance for stroke. We hypothesized that one reason why might be that the preoperative features we consider are not sufficient to predict stroke. Investigating to what extent intraoperative features predict stroke versus operative mortality could give us a better sense of what factors contribute most to these difference adverse outcomes.

- **Investigating nonlinear feature construction.** We hypothesized that another reason why the performance for stroke was inferior to the performance for operative mortality is that there may be nonlinear relationships between features that we have not explored here. Considering nonlinear feature construction may give us more insight into how preoperative features are related to the risk of stroke.

6.3 Conclusions

Making use of available auxiliary data when training models for a target task with little data can lead to significant improvements in performance. Particularly in applications where there is a high class-imbalance, transfer learning methods such as the ones discussed in this thesis could make training accurate models for very specific patient populations more feasible.

Appendix A

Table A.1: Operative Mortality: Average performance over 100 test sets of isolated AVRs for models trained on different training sets with 95% confidence intervals. These training sets combine source data from the hospital of interest, h, with data from the other hospital, $\bar{h}$. Training subsets which vary in whether or not an AVR was performed are compared.

Hospital	**Source**	**AUC**	**Precision**	**Recall**
	Data	95 % CI	95 % CI	95 % CI
1	$(S_1^{\text{AVR},2} \cup S_1^{+\text{AVR}})$	0.6575 (0.55, 0.78)	0.0317 (0.01,0.05)	0.3244 (0.11,0.56)
	$S_1^{-\text{AVR}}$	0.6843 (0.55, 0.79)	0.0450 (0.02,0.07)	0.4600 (0.22, 0.67)
	S_1	0.7179 (0.59, 0.83)	0.0462 (0.02,0.07)	0.4722 (0.22, 0.67)
2	$(S_2^{\text{AVR},1} \cup S_2^{+\text{AVR}})$	0.7371 (0.65, 0.81)	0.0770 (0.06, 0.09)	0.5474 (0.39,0.68)
	$S_2^{-\text{AVR}}$	0.7304 (0.65, 0.81)	0.0751 (0.06, 0.09)	0.5342 (0.39, 0.68)
	S_2	0.7405 (0.66, 0.81)	0.0750 (0.05, 0.10)	0.5332 (0.35, 0.70)

Table A.2: Stroke: Average performance over 100 test sets of isolated AVRs for models trained on different training sets with 95% confidence intervals. These training sets combine source data from the hospital of interest, h, with data from the other hospital, $\bar{h}$. Training subsets which vary in whether or not an AVR was performed are compared.

Hospital	**Source**	**AUC**	**Precision**	**Recall**
	Data	95 % CI	95 % CI	95 % CI
1	$(S_1^{\text{AVR}.2} \cup S_1^{+\text{AVR}})$	**0.5790 (0.45,0.68)**	**0.0188 (0,0.03**	**0.1730 (0,0.30)**
	$S_1^{-\text{AVR}}$	0.5131 (0.40,0.62)	0.0148 (0,0.03)	0.1360 (0,0.30)
	S_1	0.5468 (0.44,0.64)	0.0155 (0,0.03)	0.1430 (0,0.030)
2	$(S_2^{\text{AVR}.1} \cup S_2^{+\text{AVR}})$	0.5477 (0.47,0.65)	0.0272 (0.01,0.04)	0.1891 (0.09,0.30)
	$S_2^{-\text{AVR}}$	0.5519 (0.47,0.62)	**0.0369 (0.02,0.06)**	**0.2570 (0.13,0.39)**
	S_2	**0.5612 (0.48,0.63)**	0.0296 (0.01, 0.05)	0.2057 (0.09, 0.35)

Bibliography

[1] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[2] Gyemin Lee, Ilan Rubinfeld, and Zeeshan Syed. Adapting surgical models to individual hospitals using transfer learning. In *2012 IEEE 12th International Conference on Data Mining, Workshop on Biological Data Mining and its Applications in Healthcare (BioDM)*, pages 57–63. IEEE, 2012.

[3] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 0:1–8, 2014.

[4] Todd M Dewey, David Brown, William H Ryan, Morley A Herbert, Syma L Prince, and Michael J Mack. Reliability of risk algorithms in predicting early and late operative outcomes in high-risk patients undergoing aortic valve replacement. *The Journal of Thoracic and Cardiovascular Surgery*, 135(1):180–187, 2008.

[5] Dimitri Kalavrouziotis, Debbie Li, Karen J Buth, and Jean-Francois Légaré. The European System for Cardiac operative risk evaluation (EuroSCORE) is not appropriate for withholding surgery in high-risk patients with aortic stenosis: a retrospective cohort study. *Journal of Cardiothoracic Surgery*, 4:32, 2009.

[6] Alec Vahanian and Catherine M Otto. Risk stratification of patients with aortic stenosis. *European Heart Journal*, 31(4):416–423, 2010.

[7] Brigitte R Osswald, Vassil Gegouskov, Dominika Badowski-Zyla, Ursula Tochtermann, Gisela Thomas, Siegfried Hagl, and Eugene H Blackstone. Overestimation of aortic valve replacement risk by EuroSCORE: implications for percutaneous valve replacement. *European Heart Journal*, 30(1):74–80, 2009.

[8] Matthew Richardson, Neil Howell, Nick Freemantle, Ben Bridgewater, and Domenico Pagano. Prediction of in-hospital death following aortic valve replacement: a new accurate model. *European Journal of Cardio-Thoracic Surgery*, 43(4):704–708, 2013.

[9] Craig R Smith, Martin B Leon, Michael J Mack, D Craig Miller, Jeffrey W Moses, Lars G Svensson, E Murat Tuzcu, John G Webb, Gregory P Fontana,

Raj R Makkar, et al. Transcatheter versus surgical aortic-valve replacement in high-risk patients. *New England Journal of Medicine*, 364(23):2187–2198, 2011.

[10] Hélène Eltchaninoff, Alain Prat, Martine Gilard, Alain Leguerrier, Didier Blanchard, Gérard Fournial, Bernard Iung, Patrick Donzeau-Gouge, Christophe Tribouilloy, Jean-Louis Debrux, et al. Transcatheter aortic valve implantation: early results of the FRANCE (FRench Aortic National CoreValve and Edwards) registry. *European Heart Journal*, 32(2):191–197, 2011.

[11] Eberhard Grube, Gerhard Schuler, Lutz Buellesfeld, Ulrich Gerckens, Axel Linke, Peter Wenaweser, Barthel Sauren, Friedrich-Wilhelm Mohr, Thomas Walther, Bernfried Zickmann, et al. Percutaneous Aortic Valve Replacement for Severe Aortic Stenosis in High-Risk Patients Using the Second- and Current Third-Generation Self-Expanding CoreValve Prosthesis: Device Success and 30-day Clinical Outcome. *Journal of the American College of Cardiology*, 50(1):69–76, 2007.

[12] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To transfer or not to transfer. In *NIPS 2005 Workshop on Transfer Learning*, volume 898, 2005.

[13] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 2013.

[14] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 193–200. ACM, 2007.

[15] Jing Jiang and Chengxiang Zhai. Instance weighting for domain adaptation in NLP. In *ACL*, volume 2007, page 22, 2007.

[16] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[17] Sinno Jialin Pan, James T Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008.

[18] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128. Association for Computational Linguistics, 2006.

[19] Pengcheng Wu and Thomas G Dietterich. Improving SVM accuracy by training on auxiliary data sources. In *Proceedings of the 21st International Conference on Machine Learning*, pages 110–117. ACM, 2004.

[20] Samer AM Nashef, François Roques, Philippe Michel, E Gauducheau, S Lemeshow, R Salamon, et al. European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery*, 16(1):9–13, 1999.

[21] François Roques, Philippe Michel, Antony R Goldstone, and Samer AM Nashef. The logistic euroscore. *European Heart Journal*, 24(9):882–882, 2003.

[22] David M Shahian, Sean M O'Brien, Giovanni Filardo, Victor A Ferraris, Constance K Haan, Jeffrey B Rich, Sharon-Lise T Normand, Elizabeth R DeLong, Cynthia M Shewan, Rachel S Dokholyan, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1– coronary artery bypass grafting surgery. *The Annals of Thoracic Surgery*, 88(1):S2–S22, 2009.

[23] Sean M O'Brien, David M Shahian, Giovanni Filardo, Victor A Ferraris, Constance K Haan, Jeffrey B Rich, Sharon-Lise T Normand, Elizabeth R DeLong, Cynthia M Shewan, Rachel S Dokholyan, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2 – isolated valve surgery. *The Annals of Thoracic Surgery*, 88(1):S23–S42, 2009.

[24] David M Shahian, Sean M O'Brien, Giovanni Filardo, Victor A Ferraris, Constance K Haan, Jeffrey B Rich, Sharon-Lise T Normand, Elizabeth R DeLong, Cynthia M Shewan, Rachel S Dokholyan, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3 – valve plus coronary artery bypass grafting surgery. *The Annals of Thoracic Surgery*, 88(1):S43–S62, 2009.

[25] Samer AM Nashef, François Roques, Linda D Sharples, Johan Nilsson, Christopher Smith, Antony R Goldstone, and Ulf Lockowandt. Euroscore II. *European Journal of Cardio-Thoracic Surgery*, 41(4):734–745, 2012.

[26] Daniel Wendt, Brigitte R Osswald, Katrin Kayser, Matthias Thielmann, Paschalis Tossios, Parwis Massoudy, Markus Kamler, and Heinz Jakob. Society of Thoracic Surgeons score is superior to the EuroSCORE: determining mortality in high risk patients undergoing isolated aortic valve replacement. *The Annals of Thoracic Surgery*, 88(2):468–475, 2009.

[27] Joan Ivanov, Jack V Tu, and C David Naylor. Ready-made, recalibrated, or remodeled? issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation*, 99(16):2098–2104, 1999.

[28] Otto Pitkänen, Minna Niskanen, Sinikka Rehnberg, Mikko Hippeläinen, and Markku Hynynen. Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE. *European Journal of Cardio-Thoracic Surgery*, 18(6):703–710, 2000.

[29] Cheng-Hon Yap, Christopher Reid, Michael Yii, Michael A Rowland, Morteza Mohajeri, Peter D Skillington, Siven Seevanayagam, and Julian A Smith. Validation of the EuroSCORE model in Australia. *European Journal of Cardio-Thoracic Surgery*, 29(4):441–446, 2006.

[30] John Chalmers, Mark Pullan, Brian Fabri, James McShane, Matthew Shaw, Neeraj Mediratta, and Michael Poullis. Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. *European Journal of Cardio-Thoracic Surgery*, 43(4):688–694, 2013.

[31] G Asimakopoulos, S Al Ruzzeh, G Ambler, RZ Omar, P Punjabi, M Amrani, and KM Taylor. An evaluation of existing risk stratification models as a tool for comparison of surgical performances for coronary artery bypass grafting between institutions. *European Journal of Cardio-Thoracic Surgery*, 23(6):935–942, 2003.

[32] P Pinna-Pintor, M Bobbio, S Colangelo, F Veglia, M Giammaria, F Maisano, O Alfieri, et al. Inaccuracy of four coronary surgery risk-adjusted models to predict mortality in individual patients. *European Journal of Cardio-Thoracic Surgery*, 21(2):199–204, 2002.

[33] Johan Nilsson, Lars Algotsson, Peter Höglund, Carsten Lührs, and Johan Brandt. Comparison of 19 pre-operative risk stratification models in open-heart surgery. *European Heart Journal*, 27(7):867–874, 2006.

[34] Christopher Reid, Baki Billah, Diem Dinh, Julian Smith, Peter Skillington, Michael Yii, Seven Seevanayagam, Morteza Mohajeri, and Gil Shardey. An Australian risk prediction model for 30-day mortality after isolated coronary artery bypass: the AusScore. *The Journal of Thoracic and Cardiovascular Surgery*, 138(4):904–910, 2009.

[35] Daniel Wendt, Matthias Thielmann, Philipp Kahlert, Svea Kastner, Vivien Price, Fadi Al Rashid, Polykarpos Patsalis, Raimund Erbel, and Heinz Jakob. Comparison between different risk scoring algorithms on isolated conventional or transcatheter aortic valve replacement. *The Annals of Thoracic Surgery*, 2013.

[36] Shukri F Khuri, Jennifer Daley, William Henderson, Kwan Hur, John Demakis, J Bradley Aust, Vernon Chong, Peter J Fabri, James O Gibbs, Frederick Grover, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Annals of Surgery*, 228(4):491, 1998.

[37] Robert O Bonow, Blase A Carabello, Kanu Chatterjee, Antonio C de Leon, David P Faxon, Michael D Freed, William H Gaasch, Bruce Whitney Lytle, Rick A Nishimura, Patrick T OGara, et al. ACC/AHA 2006 Guidelines for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 1998 Guidelines for the Management of Patients With Valvular Heart Disease) Developed in Collaboration With the Society of Cardiovascular Anesthesiologists Endorsed by the Society for Cardiovascular Angiography and Interventions and the Society of Thoracic Surgeons. *Journal of the American College of Cardiology*, 48(3):e1–e148, 2006.

[38] Alain Cribier, Helene Eltchaninoff, Assaf Bash, Nicolas Borenstein, Christophe Tron, Fabrice Bauer, Genevieve Derumeaux, Frederic Anselme, François Laborde, and Martin B Leon. Percutaneous transcatheter implantation of an aortic valve prosthesis for calcific aortic stenosis first human case description. *Circulation*, 106(24):3006–3008, 2002.

[39] Food and Drug Administration. FDA approves first artificial aortic heart valve placed without open-heart surgery [Press release]. Retrieved from http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm278348.htm, 2011.

[40] Food and Drug Administration. FDA approval expands access to artificial heart valve for inoperable patients [Press release]. Retrieved from http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm369510.htm, 2013.

[41] Farah Bhatti, Antony D Grayson, Geir Grotte, Brian M Fabri, John Au, Mark Jones, and Ben Bridgewater. The logistic EuroSCORE in cardiac surgery: how well does it predict operative risk? *Heart*, 92(12):1817–1820, 2006.

[42] Ganesh Shanmugam, Mark West, and Geoff Berg. Additive and logistic EuroSCORE performance in high risk patients. *Interactive Cardiovascular and Thoracic Surgery*, 4(4):299–303, 2005.

[43] JF Gummert, A Funkat, B Osswald, A Beckmann, W Schiller, A Krian, F Beyersdorf, A Haverich, and J Cremer. EuroSCORE overestimates the risk of cardiac surgery: results from the national registry of the German Society of Thoracic and Cardiovascular Surgery. *Clinical Research in Cardiology*, 98(6):363–369, 2009.

[44] Daniel Wendt, Brigitte Osswald, Matthias Thielmann, Katrin Kayser, Paschalis Tossios, Parwis Massoudy, Markus Kamler, and Heinz Jakob. The EuroSCORE–still helpful in patients undergoing isolated aortic valve replacement? *Interactive Cardiovascular and Thoracic Surgery*, 10(2):239–244, 2010.

[45] Society of Thoracic Surgeons. STS Facts. Retrieved from http://www.sts.org/about-sts/sts-fact-sheet, 2014.

[46] Society of Thoracic Surgeons. STS Adult Cardiac Data Specifications, Version 2.41. Retrieved from http://www.sts.org/sts-national-database/database-managers/adult-cardiac-surgery-database/data-collection, 2001.

[47] Society of Thoracic Surgeons. STS Adult Cardiac Data Specifications, Version 2.52. Retrieved from http://www.sts.org/sts-national-database/database-managers/adult-cardiac-surgery-database/data-collection, 2004.

[48] Society of Thoracic Surgeons. STS Adult Cardiac Data Specifications, Version 2.61. Retrieved from http://www.sts.org/sts-national-database/database-managers/adult-cardiac-surgery-database/data-collection, 2007.

[49] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[50] Katharina Morik, Peter Brockhausen, and Thorsten Joachims. Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, 1999.

[51] Morris H DeGroot and Mark J Schervish. *Probability and Statistics.* Addison Wesley, 2002.

[52] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83, 1945.